University of Dundee

MASTER OF DENTAL SCIENCE

In-vitro Study Investigating the Validity and Reliability of a New Software for Digital Scoring of the American Board of Orthodontics Objective Grading System

El-Engebawy, Eslam M. F.

Award date:
2015

In-vitro Study Investigating the Validity and Reliability of a New Software for Digital Scoring of the American Board of Orthodontics Objective Grading System

Eslam M. F. El-Engebawy

In-vitro Study Investigating the Validity and Reliability of a New Software for Digital Scoring of the American Board of Orthodontics Objective Grading System

Eslam M. F. El-Engebawy

Submitted for the fulfilment of the requirements for the Degree of Masters of Sciences

School of Dentistry

College of Medicine, Dentistry and Nursing
University of Dundee

Feb 2015

Table of Contents

List of Figures....vi

List of Tables....vi

Acknowledgements....x

Dedication....xi

Declaration....xii

Abstract....xiii

Chapter 1: Introduction....1

Chapter 2: Literature Review....4

2.1 Diagnostic studies....5

2.1.1 Gold standards....6

2.1.2 Difficulties associated with measurements in research....6

2.2 Validity....7

2.2.1 Types of measurement validity....8

2.3 Reliability....8

2.4 Statistical methods for investigating agreement between measurements....9

2.4.1 Bland-Altman plot....9

2.5 Statistical methods for investigating correlation between measurements....10

2.6 Agreement and correlation....11

2.7 Orthodontic occlusal indices....11

2.7.1 Diagnostic indices....12

2.7.2 Epidemiologic indices....12

2.7.3 Orthodontic treatment needs indices....12

2.7.4 Orthodontic treatment complexity index....13

2.7.5 Orthodontic treatment outcome indices....14

2.7.5.1 The PAR Index (Peer Assessment Rating Index)....14

2.7.5.1.1 PAR Index advantages....15

2.7.5.2 The Index of Complexity, Outcome and Need (ICON) 15

2.7.5.3 American Board of Orthodontics Objective Grading System 16

2.7.5.3.1 American Board of Orthodontics Objective Grading System criteria and values 17

2.7.6 Comparison between treatment outcome indices 22

2.8 Digital study models 22

2.8.1 Advantages of digital study models 23

2.8.2 Construction of digital study models 23

2.8.3 Digital model software systems 25

2.8.3.1 OrthoCAD 25

2.8.3.2 emodel software 26

2.8.3.3 Other software programs 26

2.8.3.4 Comparison between OrthoCAD and emodel software 27

2.9 Research into digital model software 28

2.9.1 Accuracy of digital models 28

2.10 Studies comparing plaster and digital models measuring ABO OGS components 32

2.10.1 Sample size used in studies conducted on ABO OGS software systems 33

2.10.2 Methods used in studies conducted on ABO OGS software systems .34

2.10.3 Data analysis in studies conducted on ABO OGS software systems ..36

2.10.4. Reliability of examiners in studies conducted on ABO OGS software systems 39

2.10.5 Overall evaluation of the results in studies conducted on ABO OGS software systems 39

2.11 Time taken to compare plaster and digital models 45

2.12 The new software system 47

2.13 Summary of the literature review 48

Chapter 3: Aim and Hypotheses 50
3.1 The aim of the study 51
3.2 Null hypotheses 52
Chapter 4: Materials and Methods 53
4.1 Materials 54
4.1.1 Inclusion criteria for the study models 54
4.1.2 Scanning and digitisation of the models 54
4.1.3 The ABO Gauge (Casko et al., 1998) 56
4.1.4 Examiners involved in the study 56
4.1.5 The new software system 57
4.2 Methods 57
4.2.1 Examiners' training and calibration 57
4.2.2 Examiners' scoring 58
4.2.3 Software advantages 59
4.3 Statistical analysis 60
4.3.1 Descriptive analysis 60
4.3.2 Examiner reliability 60
4.3.3 Agreement between methods 60
4.4 Time 61
4.4.1 Methods 61
4.4.2 Time statistical analysis between the two methods 61
Chapter 5: Results 62
5.1 Descriptive analysis 63
5.1.1 Mean, standard deviation and range for plaster models 63
5.1.2 Mean, standard deviation and range for digital models 64
5.1.3 The difference in points between digital and plaster models 65
5.2 Examiner reliability 70

5.2.1 Examiners' reliability for plaster models 71

5.2.2 Examiners' reliability for digital models 72

5.3 Validation of the new software 74

5.4 Distribution of the scoring result of the ABO OGS for the sample using the conventional plaster model and digital models 89

5.5 Time for ABO scoring 90

5.5.1 Time for ABO OGS scoring digital models using the new software 90

5.5.2 Time for ABO scoring using plaster models 91

5.5.3 Comparison between times for ABO OGS scoring using plaster and digital models 91

Chapter 6: Discussion 93

6.1 Descriptive analysis 94

6.1.1 Study sample size 94

6.1.2 Description of the ABO OGS scores 94

6.2 Reliability 97

6.2.1 Intra-examiner reliability 97

6.2.1.1 Plaster models 97

6.2.1.2 Digital models 98

6.2.1.3 Intra-examiner reliability comparison with previous studies 99

6.2.2 Inter-examiner reliability 99

6.2.2.1 Plaster models 100

6.2.2.2 Digital models 101

6.2.2.3 Inter-examiner reliability comparison with previous studies 101

6.3 Validity of the new software system 103

6.3.1 Statistical analysis used in comparing plaster and digital models 103

6.3.2 Validity and accuracy of the new software 104

6.3.3 Comparison of current results with previous studies 105

6.4 Time for ABO OGS scoring 107

6.4.1 Comparison among examiners for scoring time 108

6.4.1.1 Time scoring plaster models 108

6.4.1.2 Time scoring digital models 109

6.5 Sources of error in the ABO software system 109

6.6 Clinical implications 109

Chapter 7: Conclusion 111

Appendix 1: Bland-Altman plots and tables for plaster and digital models 120

Appendix 2: Grading System for Dental Casts and Panoramic Radiographs 138

List of Figures

Figure 4.1 ABO gauge (Casko et al., 1998).....56
Figure 4.2 Screenshot of a digital model using the new software 59
Figure 4.3 Rotating the digital model using the new software 59
Figure 5.1 Bland-Altman scatter plot for alignment for Examiner1.....74
Figure 5.2 Bland-Altman scatter plot for the marginal ridge for Examiner 1.....75
Figure 5.3 Bland-Altman scatter plot for the buccolingual inclination for Examiner 1 ..76
Figure 5.4 Bland-Altman scatter plot for the buccolingual inclination for Examiner 2 ..77
Figure 5.5 Bland-Altman scatter plot for the buccolingual inclination for Examiner 3 ..78
Figure 5.6 Bland-Altman scatter plot for the buccolingual inclination for Examiner 4 ..79
Figure 5.7 Bland-Altman scatter plot for the occlusal relationship for Examiner 1.....80
Figure 5.8 Bland-Altman scatter plot for the occlusal contacts for Examiner 1.....81
Figure 5.9 Bland-Altman scatter plot for the overjet for Examiner 1.....82
Figure 5.10 Bland-Altman scatter plot for the inter-proximal contacts for Examiner 1..83
Figure 5.11 Bland-Altman scatter plot for the total score for Examiner 1 84
Figure 5.12 Bland-Altman scatter plot for the total score for Examiner 2 85
Figure 5.13 Bland-Altman scatter plot for the total score for Examiner 3 86
Figure 5.14 Bland-Altman scatter plot for the total score for Examiner 4 87
Figure 5.15 Bland-Altman scatter plot for Alignment for Examiner 2.....120
Figure 5.16 Bland-Altman scatter plot for Marginal Ridges for Examiner 2 121
Figure 5.17 Bland-Altman scatter plot for Occlusal Relationship for Examiner 2.....122
Figure 5.18 Bland-Altman scatter plot for Occlusal Contacts for Examiner 2.....123
Figure 5.19 Bland-Altman scatter plot for Overjet for Examiner 2.....124
Figure 5.20 Bland-Altman scatter plot for Interproximal Contacts for Examiner 2.....125
Figure 5.21 Bland-Altman scatter plot for Alignment for Examiner 3.....126
Figure 5.22 Bland-Altman scatter plot for Marginal Ridges for Examiner 3 127
Figure 5.23 Bland-Altman scatter plot for Occlusal Relationship of Examiner 3.....128
Figure 5.24 Bland-Altman scatter plot for Occlusal Contact for Examiner 3 129
Figure 5.25 Bland-Altman scatter plot for Overjet for Examiner 3.....130
Figure 5.26 Bland-Altman scatter plot for Interproximal Contacts for Examiner 3.....131
Figure 5.27 Bland-Altman scatter plot for Alignment for Examiner 4.....132
Figure 5.28 Bland-Altman scatter plot for Marginal Ridge for Examiner 4.....133
Figure 5.29 Bland-Altman scatter plot for Occlusal Relationship for Examiner 4.....134
Figure 5.30 Bland- Altman scatter plot for Occlusal Contacts for Examiner 4.....135

Figure 5.31 Bland-Altman scatter plot for Overjet for Examiner 4.....136
Figure 5.32 Bland-Altman scatter plot for Interproximal Contacts for Examiner 4.....137

List of Tables

Table 2.1 ABO Cast/ Radiograph Evaluation 20
Table 2.2 Comparison between OrthoCAD and emodel software 27
Table 2.3 Studies that investigated different components of the ABO score.....38
Table 2.4 Studies that investigated different components of the ABO score.....44
Table 5.1 Mean, standard deviation and range for the four examiners when measuring the seven components of the ABO OGS and total scores using 31 plaster models 67
Table 5 .2 Mean, standard deviation and range for the four examiners when measuring the seven components of the ABO OGS and total scores using 31 digital models 68
Table 5 .3Mean, standard deviation and range of the differences between plaster and digital for the four examiners.....69
Table 5.4 Pearson correlation coefficients for comparisons of repeated measurements for Examiner 1 of plaster and digital models.....70
Table 5.5 Pearson Correlation coefficients and P values for comparisons between each examiner and record type for each ABO OGS component and total score to show inter-examiner reliability for plaster models.....71
Table 5.6 Pearson correlation coefficients and P values for comparisons between each examiner and record type for each ABO OGS component and total score to show inter-examiner reliability for digital models. 72
Table 5.7 Mean difference and limit of agreements of plaster and digital models measurements 88
Table 5.8 distribution of the scoring result of the ABO OGS (Examiners 1-4) for the sample using the conventional plaster model and the ABO software 89
Table 5. 9 ANOVA test to compare the time taken to score thirty one digital models.....90
Table 5.10 ANOVA test to compare the time taken by different examiners to score the plaster models.....91
Table 5.11 Paired t-test to compare the time taken for ABO OGS scoring between the plaster and digital models 92
Table 5.12 Mean and limits of agreement of plaster and digital models for Alignment for Examiner 2 120

Table 5.13 Mean and limits of agreement of plaster and digital models for Marginal Ridge for Examiner 2 121
Table 5.14 Mean and limits of agreement of plaster and digital models for Occlusal Relationship for Examiner 2 122
Table 5.15 Mean and limits of agreement of plaster and digital models for Occlusal Contacts for Examiner 2 123
Table 5.16 Mean and limits of agreement of plaster and digital models for Overjet for Examiner 2 124
Table 5.17 Mean and limits of agreement of plaster and digital models for Interproximal Contacts for Examiner 2 125
Table 5.18 Mean and limits of agreement of plaster and digital models for Alignment for Examiner 3 126
Table 5.19 Mean and limits of agreement of plaster and digital models for Marginal Ridge for Examiner 3 127
Table 5. 20 Mean and limits of agreement of plaster and digital models for Occlusal Relationship for Examiner 3 128
Table 5.21 Mean and limits of agreement of plaster and digital models for Occlusal Contacts for Examiner 3 129
Table 5.22 Mean and limits of agreement of plaster and digital models for Overjet for Examiner 3 130
Table 5.23 Mean and limits of agreement of plaster and digital models for Interproximal Contacts for Examiner 3 131
Table 5.24 Mean and limits of agreement of plaster and digital models for Alignment for Examiner 4 132
Table 5.25 Mean and limits of agreement of plaster and digital models for Marginal Ridge for Examiner 4 133
Table 5.26 Mean and limits of agreement of plaster and digital models for Occlusal Relationship for Examiner 4 134
Table 5.27 Mean and limits of agreement of plaster and digital models for Occlusal Contacts for Examiner 4 135
Table 5.28 Mean and limits of agreement of plaster and digital models for Overjet for Examiner 4 136
Table 5.29 Mean and limits of agreement of plaster and digital models for Interproximal Contacts for Examiner 4 137

Acknowledgements

Thanks and gratitude should go first and foremost to Allah, the Almighty God who endowed me with the energy to complete this thesis.

I am truly and deeply indebted to Professor Mark Hector, Dean of Dental School and Head of Department for believing in me and giving me the opportunity to succeed in my career.

I am particularly indebted to my supervisor, Professor David Bearn for his support, encouragement and guidance throughout this thesis. I am very thankful for his patience during the reading and correction of this thesis and I am very grateful for his expert advice. I owe my deepest gratitude to my supervisor, Professor Peter Mossey for his patience, motivation, enthusiasm, guidance and feedback, valuable time and effort to support the completion of this thesis. I could not have imagined having better supervisors for my master's study.

Special thanks to Dr Ahmed El-Angbawi, Dr Nasr and Dr Colin Ritche. Their continuous advice, help and immense knowledge was of great value in this thesis.

I would like to express my special gratitude and thanks to the administrative of Orthodontic Department for helping in organizing the models collection.

I would also like to thank my beloved father and mother who always supported me spiritually throughout my life and throughout writing this thesis.

Dedication

I would like to dedicate this thesis to my dear beloved parents.

Declaration

I declare that the work presented in this thesis is all my own work, has not previously been accepted for a higher degree and I have consulted all references cited.

Signed: E Slam

Date: 5/3/2015

I declare that the conditions of the relevant Ordinance and Regulations have been fulfilled.

Signed:

Date:

Abstract

Aim: The primary aim of this study is to investigate the validity and reliability of new software in order to score the American Board of Orthodontics Objective Grading System

(ABO OGS) by using digital models to replace the conventional method (gold standard) of using plaster models and a hand ABO gauge. The secondary aims of the study are to assess the level of agreement between orthodontists and non-orthodontists' ABO OGS scores using the conventional method on plaster models and the new software on digital models and then compare the time taken to score the ABO OGS using both methods.

Design: In-vitro study

Materials and methods: Thirty one high quality post-treatment plaster models which met the agreed inclusion criteria were used in this study. All models were scanned in order to compose digital models using a laser scanner (3 shape Scan IT Orthodontics). Four examiners with different levels of orthodontic knowledge (orthodontic postgraduate student, orthodontist, prosthodontic postgraduate student and undergraduate dental student) participated in the study. The four examiners scored the ABO OGS for the 31 models using the conventional method on plaster models and on digital models using the new software. All examiners received the appropriate training for using both ABO OGS scoring systems prior to scoring the models.

To determine the intra-examiner reliability, Examiner 1 repeated the ABO OGS scoring for the 7 components twice with a two-week interval. To determine the inter-examiner reliability, the 4 examiners measured the 31 models using both

scoring methods. The scores of all 4 examiners for the 31 models were used to assess the correlation and agreement between the new software and the conventional method (gold standard) for the ABO OGS' seven components (tooth alignment, vertical positioning of marginal ridges, buccolingual inclination of posterior teeth, occlusal relationship, occlusal contacts, overjet and interproximal contacts). The time taken by the 4 examiners for scoring was compared between the new software and the conventional method (gold standard).

The Pearson correlation coefficients test and Bland-Altman plots were used to assess the correlation and agreement between the two scoring methods as well as the intra and inter-examiners reliability. An ANOVA test and a t-test were used to assess the difference in scoring time for the mean of the 31 models.

Primary outcomes: The agreement between the new software and the conventional method (gold standard) for scoring ABO OGS. The intra-examiner and inter-examiner reliability of both scoring methods.

Secondary outcomes: Time taken for scoring the ABO OGS between the two methods and between examiners with different orthodontic knowledge levels.

Results: There was no agreement between the new software and the conventional scoring method in any of the ABO OGS components, except in regard to the buccolingual inclination component.

The new software had acceptable intra-examiner reliability: the total score was R = 0.583 (P = 0.001); however, inter-examiner reliability was found to be low, with a correlation of R = -0.039 (P = 0.834). The conventional scoring method had high intra-examiner and inter-examiner reliability, with correlation as high as

R =0.915 (P = 0.000) for intra-examiner reliability and R = 0.970 (P=P = 0.000) for inter-examiner reliability for the total score. Differences in the level of orthodontic knowledge among examiners had a significant influence on the reliability of the new software. However, this influence was not found to be significant in the conventional scoring method.

The new software took significantly more time for ABO OGS scoring than the conventional plaster model method (mean difference 20.09 minutes, P = 0.00).

Conclusion: The new software is not a valid method for ABO OGS scoring; it cannot replace the conventional plaster model. The new software requires more time for scoring.

Chapter 1: Introduction

The Objective Grading System (OGS) was developed by the American Board of Orthodontics (ABO) and is commonly used for assessing treatment outcomes. It is an index that can determine the success of orthodontic treatment using post-treatment dental models and panoramic radiographs.

There are seven components of the ABO OGS: alignment, marginal ridge levelling, buccolingual inclination, occlusal contacts, occlusal relationship, overjet and interproximal contacts. However, the panoramic radiograph component has a single element, which is root angulation (Appendix 2). To achieve an acceptable score on the ABO examination, more than 30 points deduction from the total maximum score will be considered as failure (with the total score consisting of the sum of the seven ABO OGS components), while 20 points may be missed in order to pass the examination (Appendix 2). Orthodontic study models are essential tools in treatment planning and as a record of treatment progress and outcome. However, due to a lack of space, the storage of study models can cause fractures. Digital models have been used in an attempt to solve this problem of the long-term storage of orthodontic study models. Digital models have benefits including inexpensive storage, easy retrieval, facility of duplication and transmission around the world.

As a result of the rapid development in computer software and three-dimensional technologies, it has become possible to create software that measures the various dimensions required to calculate indices.

In the current study, it was decided to use a newly developed piece of software to calculate ABO OGS scores for digital models. It is important to stress that the software in the current study was not developed by the American Board of Orthodontics. The new software scoring will be compared with conventional scoring on plaster models to determine whether there is agreement between the

two methods, to assess whether it is as accurate and reliable as the conventional method, and to evaluate whether there is an acceptable difference in time to score, so that we can decide if using the digital models and new software can replace conventional scoring on plaster.

Chapter 2: Literature Review

2.1 Diagnostic studies

Diagnostic accuracy studies are an important step in the evaluation of new diagnostic technologies. Diagnostic accuracy studies aim to measure the amount of agreement between index test results (new test) and the outcome of the reference standard (or gold standard). The term "accuracy" has been defined as the closeness of agreement between an analytical measurement and its actual value (Bossuyt et al., 2003)**.** Several factors threaten the internal and external validity of a study of diagnostic accuracy. Some of these factors concern the design of such studies, others relate to the selection of patients, and other factors include the execution of the tests or the analysis of the data. When evaluating a diagnostic accuracy study, it is, therefore, essential to consider both the potential for bias and sources of variation, which determine applicability (Whiting et al., 2004)**.**

Use of the terms "sensitivity" and "specificity" to describe a test is common in diagnostic studies. These terms can describe the accuracy or validity of the test used and can influence decisions related to starting, stopping or modifying treatment. However, the practical value of a diagnostic test can only be assessed by taking into account subsequent health outcomes (Mol et al., 2003).

A recent systematic review designed to evaluate the sources of bias and variation in diagnostic studies reported that there was consistent evidence that demographic features influence sensitivity, and that higher disease prevalence and severity, experienced observers, availability of clinical information, inappropriate reference standards, and partial and differential verification bias increase estimates of sensitivity and/or overall accuracy (Whiting et al., 2013).There was some evidence that retrospective data collection, data-driven threshold selection, improvements in test technology, differences in test

execution, and knowledge of the index test results when interpreting the reference standard increase sensitivity. Effects on specificity were less clear. Evidence was limited for other sources of bias and variation (Whiting et al., 2013).

2.1.1 Gold standards

In medicine, a "gold standard test" refers to a diagnostic test that is the best one available under reasonable conditions. Sometimes a "gold standard test" refers to the best-performing test available (reference test). However, although a test may not be ideal, it may still be set as the gold standard because it is the best-performing test available. In diagnostic studies, the aim is to measure the amount of agreement between index test results and the outcome of the reference gold standard.

2.1.2 Difficulties associated with measurements in research

All measurements are subject to uncertainty, and a measured value is only complete if it is accompanied by a statement of the associated uncertainty. When a quantity is measured, the outcome depends on the measuring system, the measurement procedure, the skill of the operator, the environment and other effects (Bell, 2001)**.** It is important to understand the factors that contribute to measurement errors in order to determine the appropriate actions that should be taken to improve accuracy. Measurement errors are classified into two categories:

- A systematic error (an estimate of which is known as a measurement bias) is associated with the fact that a measured value contains an offset. In general, a systematic error, regarded as a quantity, is a component of error that remains constant or depends in a specific manner on some other quantity.

- A random error is associated with the fact that, when a measurement is repeated, it will generally provide a measured value that is different from the previous value. It is random in that the next measured value cannot be predicted exactly from such previous values.

2.2 Validity

Validity of a measurement refers to the degree to which a measurement accurately reflects or assesses the specific concept that the researcher is attempting to measure. Researchers should be concerned with both *external* and *internal* validity. External validity refers to the extent to which the results of a study are generalisable or transferable. Internal validity refers to the rigour with which the study was conducted (e.g. the study design, the care taken to conduct measurements, and decisions concerning what was and was not measured) and the extent to which the designers of a study have taken into account alternative explanations for any causal relationships they explore (Peat, 2002).

Whenever a test or other measuring tool is used as part of the data collection process the validity and reliability of that test is important. Validity of a certain diagnostic test/index is a measure of how accurately it can measure what it is intended to measure. This can be a process of gathering evidence to provide a scientific base for interpreting the scores of a test/index, which may sometimes lead to setting a “gold standard” test/index to compare with the test/index being investigated (Messick, 1989).

2.2.1 Types of measurement validity

1. Several traditional validity types have been described (Messick, 1989): Content validity is evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn.
2. Criterion-related validity is evaluated by comparing the test scores with one or more external variables (called criteria) considered in order to provide a direct measure of the characteristic or behaviour in question.
3. Predictive validity indicates the extent to which an individual's future evaluation on the criterion is predicted using prior test performance.
4. Concurrent validity indicates the extent to which the test scores estimate an individual's present standing on the criterion.
5. Construct validity is evaluated by investigating the qualities measured by a test and by determining the degree to which certain explanatory concepts or constructs account for performance.

In the current study, construct validity was used to assess the validity of a new piece of software that was constructed to score the ABO OGS against the conventional method using plaster models.

2.3 Reliability

Reliability is the extent to which an experiment, test or any measuring procedure yields the same result in repeated trials. Without the agreement of independent observers who are able to replicate research procedures, or the ability to use research tools and procedures that yield consistent measurements, researchers would be unable to satisfactorily draw conclusions, formulate theories or make claims about the generalisability of their research (Allen and Yen, 2001).

There are four general classes of reliability estimates, each of which assesses reliability in a different way:

1. Inter-rater or inter-observer reliability: used to assess the degree to which different raters/observers give consistent estimates of the same phenomenon.
2. Test-retest reliability: used to assess the consistency of a measure from one time point to another.
3. Parallel-forms reliability: used to assess the consistency of the results of two tests constructed in the same way from the same content domain.
4. Internal consistency reliability: used to assess the consistency of results across items within a test.

Assessment of content and construct validity begin with non-statistical approaches including peer and/or expert review and field-testing with debriefing. Using concurrent independent measures, new tools may be compared (correlated) with established tools that are recognised as valid in a particular area.

2.4 Statistical methods for investigating agreement between measurements

2.4.1 Bland-Altman plot

Bland-Altman plots are used to assess the agreement between two measurement techniques and can also investigate the existence of any systematic difference between measurements. The mean difference is the estimated bias, and the standard deviation of the differences measures the random fluctuations around this mean. If the mean value of the difference differs significantly from zero, this indicates the presence of systematic bias. It is common to apply 95% limits of agreement for each comparison (average difference ±1.96 standard deviation

of the difference), which tells us how far apart measurements by two methods were likely to be for most individuals. If the differences within a mean of ±1.96 SD are not clinically important, the two methods may be used interchangeably (Bland and Altman, 1999, 1986). Bland-Altman plots can compare two clinical measurement techniques. They can also be used to compare a new measurement technique or method against a gold standard. In addition, the Bland-Altman plot may also be used to assess the *repeatability* of a method by comparing repeated measurements using a single method on a series of subjects. The graph can then also be used to check whether the variability or precision of a method is related to the size of the characteristic being measured (Hanneman, 2008).

2.5 Statistical methods for investigating correlation between measurements

Correlation between variables is a measure of how well the variables are related. The most common measure of correlation in statistics is the Pearson correlation, which shows the linear relationship between two variables. Two letters are used to represent the Pearson correlation: the Greek letter rho (ρ) for a population and the letter "r" for a sample. The correlation coefficient, denoted by r, is a measure of the strength of the straight line or linear relationship between two variables. The correlation coefficient assumes values ranging between +1 and -1. Positive values indicate that the two variables are positively correlated, meaning the two variables vary in the same direction. Negative values indicate that the two variables are negatively correlated, meaning that the two variables vary in the contrary direction. Values close to +1 or -1 indicate that the two variables are highly related (Hanneman, 2008).

2.6 Agreement and correlation

Correlation measures the strength of a relation between two variables, not the level of agreement between them. Two variables may have perfect agreement only if the points plotted on a graph lie along the line of equality, but will have perfect correlation if the points lie along any straight line. Bland and Altman make the point that any two methods that are designed to measure the same parameter should have good correlation when a set of samples is chosen such that the parameter to be determined varies considerably. A high correlation for any two methods designed to measure the same property could thus in itself just be a sign that one has chosen a widely spread sample. A high correlation does not automatically imply that there is good agreement between the two methods (Bland and Altman, 1986).

2.7 Orthodontic occlusal indices

The main reason for undertaking orthodontic treatment is to augment oral health, to improve functioning of the dentition, and to enhance facial and dental aesthetics (Borzabadi-Farahani and Borzabadi-Farahani, 2011). Several occlusal indices have been developed to help assess success in achieving these aims.

Occlusal indices also allow communication between orthodontists, allow estimation of the prevalence of malocclusion in a given population or individual patient, determine the level of treatment need, or assess the amount of deviation from normal occlusion. Occlusal indices can also evaluate the outcome of patient treatment and the changes resulting from treatment, and they can thus facilitate the evaluation of treatment success. Occlusal indices can, in addition, show the severity of the patient case and the complexity of the treatment required, which helps the orthodontist (Daniels and Richmond, 2000,

Shaw et al., 1995).

Occlusal indices should be valid and reliable in order to be clinically acceptable (Carlos, 1970). Many occlusal indices are validated by comparing the scores with an accepted gold standard, or commonly, the subjective consensus opinion of a group of experienced specialists (Richmond et al., 1992a). A large number of occlusal indices began to appear in the 1950s and 1960s, and these can be divided into five different categories, which are diagnostic, epidemiologic, orthodontic treatment need, treatment outcome and treatment complexity indices (Younis et al., 1997, Shaw et al., 1995).

2.7.1 Diagnostic indices

The best known is Angle's classification system (Angle, 1899), which aids communication between orthodontists by classifying malocclusion into three classes with divisions to describe it.

Other examples of diagnostic indices are the incisal categories of Ballard and Wayman (Ballard and Wayman, 1965) and the five-point system of Ackerman and Proffit (Ackerman and Proffit, 1969).

2.7.2 Epidemiologic indices

These indices record the characteristics of malocclusion in order to allow its estimation in a population. Such indices include Summers' Occlusal Index, which identifies a scoring occlusal disorder (Summers, 1971), epidemiologic registration via the Björk method (Björk et al., 1964), and Little's irregularity index (LII) (Little, 1975) for irregularity and teeth alignment.

2.7.3 Orthodontic treatment needs indices

This type of index categorises malocclusion based on treatment need, as malocclusion is not an acquired condition, but a developmental condition characterised by a deviation from normal.

Usually, these indices summarise the malocclusion with a numeric value. Treatment need indices yield a score for each trait or component, which is then weighted to calculate an overall score. Sometimes, cut-off points are used to identify those patients who do or do not need orthodontic treatment.

Examples of indices of treatment need are:

- Handicapping Labio-Lin Deviation Index (HLD) (Draker, 1960, Draker, 1967).
- Swedish Medical Board Index SMBI (Swedish Medical Health Board, 1967);(Linder-Aronson, 1974).
- Dental Aesthetic Index DAI (Cons et al., 1986).
- Summers' occlusal index SOI (Summers, 1971).
- Index of Orthodontic Treatment Need IOTN (Brook and Shaw, 1989).
- Index of Complexity, Outcome and Need ICON (Daniels and Richmond ,2000)

The Index of Orthodontic Treatment Need (IOTN) is the most commonly used occlusal index, and it comprises two components: the Aesthetic Component (AC) and the Dental Heath Component (DHC).

The AC consists of ten photographs showing different levels of dental attractiveness on a scale of 1 to 10. A score of 1 is the most attractive and a score of 10 the least attractive arrangement of teeth. The DHC categorises the detrimental effects of the various deviant occlusal traits in order of severity.

2.7.4 Orthodontic treatment complexity index

Assessing the complexity of malocclusion helps to inform the patient of likely treatment success, identify the most proper setting in which the patient receives treatment and identify cases that are more difficult and are likely to take longer

to treat (Llewellyn et al., 2007).

Examples of these indices include:

- Index of Orthodontic Treatment Complexity (IOTC) (Llewellyn et al., 2007).
- Index of Complexity, Outcome and Need (ICON) (Daniels and Richmond, 2000).

2.7.5 Orthodontic treatment outcome indices

2.7.5.1 The PAR Index (Peer Assessment Rating Index)

The PAR Index is one of the most common indices used to measure treatment outcomes. It was developed to provide a single summary score for all of the occlusal anomalies that can be found in a malocclusion by measuring the pre-treatment and post-treatment models. The difference in scores between pre-treatment and post-treatment shows the improvement in the occlusion as a result of treatment. The higher the PAR score, the greater the deviation from ideal occlusion (Richmond et al., 1992b).

In 1987, a group of ten experienced orthodontists (the British Orthodontic Standards Working Party) met and used four projection screens to see all of the images from each case simultaneously; the data was then entered into a computer database to make it easier and faster for comparison until they reached agreement. The group used more than two hundred models. The index was validated using this assessment of deviation from normal occlusion as the gold standard (Richmond et al., 1992b).

There are five components in the PAR Index (Richmond et al., 1992b):

1. Left and right buccal occlusion assessment in the anteroposterior, vertical and transverse dimensions.
2. Upper and lower anterior segments assessed for impacted teeth

alignment and spacing.

3. Overbite and open-bite for all incisors.
4. Overjet recording the number of incisors in anterior crossbite or the increased overjet of all incisors.
5. Centreline in respect of the lower midline is recorded.

The PAR Index scores were then weighted for each component in order to produce a total weighted score. Three grades of improvement were developed – "Greatly improved", "Improved" and "Worse or no difference". The weighting cases show a real effect and agreement with indices estimating treatment need (Hamdan and Rock, 1999).

2.7.5.1.1 PAR Index advantages

The PAR Index has several advantages when compared with other indices:

1. It was specially designed to provide a more objective assessment of treatment outcomes (Richmond et al., 1992b)
2. It is extensively used as a method of outcome assessment in Europe.
3. It was validated using assessment deviation from normal occlusion as the gold standard (Daniels and Richmond, 2000).
4. The index measures the deviation of teeth from an ideal position or from normal occlusion (Richmond et al., 1992a).

2.7.5.2 The Index of Complexity, Outcome and Need (ICON)

In 2000, Daniels and Richmond developed an Index of Outcome, Complexity and Need (ICON) in order to assess treatment need, outcome and case complexity. The opinions of 97 orthodontists from eight European countries and the United States of America were used to develop the ICON index. It is a multifunctional index assessing treatment need, treatment outcome, and treatment complexity (Daniels and Richmond, 2000).

The ICON consists of five components:

1. The Aesthetic Component (AC), which is similar to the AC of the IOTN (Brook and Shaw, 1989).
2. Upper arch crowding/spacing assessment.
3. Crossbite.
4. Anterior vertical relationship.
5. Buccal segment antero-posterior relationship.

The occlusal anomalies are scored, and weighted scores are subsequently summed up to produce the final ICON score.

A score of 30 or less indicates that end treatment occlusion is acceptable in assessing the treatment outcome. The outcome of orthodontic treatment is assessed by using the improvement grade, with grade assessments of "greatly improved", "substantially improved", "moderately improved", "minimally improved", "not improved" or "worse". Each component can be measured on patients as well as on study models.

Each case requires approximately one minute for the practical application of the test, which indicates that it is simple to use. The ICON was found to be valid for assessing the complexity and outcome of orthodontics cases (Daniels and Richmond, 2000).

2.7.5.3 American Board of Orthodontics Objective Grading System

The Objective Grading System (OGS) for dental casts and panoramic radiographs was introduced by the American Board of Orthodontics (ABO) in 1998. The ABO was aiming to make postgraduate clinical examinations objective, accurate and reliable with the goal of maintaining the highest standards of clinical excellence. They developed the OGS over a period of five years, with a series of four field tests starting in 1995.

In 1995, the first trial of clinical examination on dental casts and panoramic radiographs was undertaken, with 100 cases being evaluated over a series of 15 criteria. It was found that over 85% of inadequacies in the final outcome of orthodontic treatment occurred in seven of the 15 criteria (alignment, marginal ridges, buccolingual inclination, overjet, occlusal relationships, occlusal contacts and root angulation) (Casko et al., 1998).

A second field test was conducted in 1996 to test the reliability and verify the result of the test that was previously performed. By using 300 post-treatment models and panoramic radiographs, they found that the majority of the inadequacies in the results were as the same seven components established previously; however, they had difficulty with inter-examiner reliability, so they agreed to develop an instrument for the purpose of making the measurements more reliable.

A total of 832 dental casts and all of the directors were involved in a further test, carried out in 1997. They used an instrument to measure the same seven criteria more accurately and a calibration was performed during that test to improve reliability and establish accuracy when using the instrument (shown in Figure 4.1), as seen in the Materials and Methods chapter. As a result, the directors decided to make some modifications to the instrument to improve measurement accuracy: one more component, interproximal contacts, was added, so that the ABO OGS would have eight components instead of seven. The final test was conducted in 1998, and all directors participated. Calibrations were performed with the modified instrument.

2.7.5.3.1 American Board of Orthodontics Objective Grading System criteria and values

In general, cases start with a presuming score of zero and for each imperfection the detected points are deducted. As a rule, cases which lose more than 30

points in the overall grading would fail the American Board of Orthodontics exam and cases that are deducted less than 20 points would pass.

The maximum score for alignment is 64 points. The scoring mechanism deducts 1 point for each contact point alignment by 0.5 and 1 mm, and if the misalignment is greater than 1 mm, 2 points are subtracted.

The marginal ridge should be at the same level or within a 0.5 mm irregularity. If the ridges are unlevelled by 0.5 to 1 mm, 1 point is subtracted from the proximal contacts. If the discrepancy is greater than 1 mm, 2 points are subtracted. The maximum points that can be subtracted are 32 from the marginal ridge score.

For the buccolingual component, the maximum points that can be deducted are 40, and if the discrepancies are between 1 and 2 mm, 1 point can be deducted for the tooth. If the discrepancy is more than 2 mm, the maximum deduction is 2 points.

The occlusal relationship should fall within 1 mm. If the deviation is between 1 and 2 mm, 1 point is subtracted per tooth. If the deviation is greater than 2 mm, 2 points is the maximum deduction per tooth. The sum is then deducted from the total occlusal relationship score.

For the occlusal contacts, if the deviation is less than 1 mm and the posterior cusp is not contacting the opposing arch, 1 point is deducted. If the distance is greater than 1 mm, 2 points are deducted. The total deductions are subtracted from 64 points, which is the maximum for the occlusal contact component.

For overjet, if there is a distance between the buccal lower cusp and the central occlusal surface, 1 point is deducted if it is less than 1 mm, and the same score is assigned for the anterior overjet. If there is more than 1 mm, 2 points will be deducted. The maximum score of overjet is 32 points.

The interproximal contacts are made from the occlusal view of the models. If there is no interproximal contact found between two teeth, and the space is up to 1 mm, 1 point is deducted for the contact between the two teeth. If greater than 1 mm, then two points will be deducted. The sum of deducted points of the interproximal contacts component will be subtracted from 60 points, which is the maximum.

The seven ABOOGS components have a combined total of 340 points. Table 2.1 explains how the ABO OGS scoring is calculated.

Table 2.1 ABO Cast/ Radiograph Evaluation

ABO Cast/Radiograph Evaluation	
ALIGNMENT/ROTATIONS 0.5mm-1mm= 1 for each tooth > 1mm = 2 for each tooth	**OCCLUSAL CONTACTS ***** 0mm = satisfactory (for each ≤1mm = 1 posterior tooth out of 1mm = 2 contact) ***Do **not** score diminutive distolingual cusps of the maxillary 1st and 2nd molars, nor lingual cusps of the mandibular first premolars. **Maximum of 2 points per tooth.**
MARGINAL RIDGES 0.5-1mm =1 >1mm= 2 (for each interproximal contact between posterior teeth) Do not include the canine-premolar contact	**OCCLUSAL RELATIONSHIP** < 1mm= satisfactory 1 – 2mm = 1(for each maxillary tooth from the canines to the 2nd molars) > 2mm = 2

Do not include the distal lower 1st premolar	
BUCCOLINGUAL INCLINATION ** 0-1mm= satisfactory 1.1 – 2mm =1 (for each posterior tooth) >2mm = 2 Do not score the mandibular 1st premolars nor the distal cusps if the second molars	**INTERPROXIMAL CONTACTS** 0.6 – 1mm= 1 >1= 2 (for each tooth interproximal contact)
OVERJET 0mm= Satisfactory <= 1mm= 1 >1mm= 2 (for each maxillary tooth)	**ROOT ANGULATION** Parallel=0 Not Parallel= 1 Root contacting Adjacent root= 2 (for each occurrence) Do not score the maxillary and mandibular canines
NOTE: Gauge Width is 0.5 mm; Gauge Height is 1 mm Third molars are not scored unless the substitute for the second molars.	

2.7.6 Comparison between treatment outcome indices

The ABO OGS detects minor variations in tooth position, which is an important factor found when James (2002) compared it with the PAR Index, but it defines only the treatment outcome and cannot take into account the severity of initial malocclusion or difficulty of treatment, which are variables that affect treatment goals. In contrast, the PAR Index scores the pre- and post-treatment dental casts, and the difference between the score represents the degree of improvement as a result of the treatment (Richmond et al., 1992a). The PAR Index may not be precise enough to discern between excellent and good final occlusion (Firestone et al., 2002) and does not evaluate periodontal health, root resorption or tooth angulation.

Onyeaso and Begole (2007) reported that PAR and ABO-OGS demonstrated 20 significant correlations with ICON in relation to treatment outcome. The authors concluded that ICON can be used in place of PAR and ABO-OGS to assess treatment outcome. The ICON index is a multifunctional index that assesses treatment need, treatment outcome, and treatment complexity.

2.8 Digital study models

The fast and continuous advances in computer sciences have resulted in the increased usage of new technologies on all levels of modern society. Digital technology is expanding within various scientific fields and is now an integral component in the field of orthodontics, in which three-dimensional imaging and modelling have undergone significant advances in recent years. As a result, soft tissues, teeth, bone and plaster models can be recreated as three-dimensional images (Hajeer et al., 2004).

Study models are very important for orthodontists because they are fundamental to diagnosis, treatment planning and evaluation of the treatment

progress and results, and are therefore a standard component of orthodontics records. Digital study models offer orthodontists an alternative to the routinely used plaster models (Santoro et al., 2003).

2.8.1 Advantages of digital study models

The advantages of digital models (Mayers et al., 2005) may include:

- Images may be transferred anywhere in the world for instant referral or consultation.
- Instant accessibility of 3D information without need for the retrieval of plaster models from storage.
- Easy storage.
- Improved quality and efficiency.
- The ability to perform accurate and simple diagnostic set-ups of various treatment options.

Plaster models are associated with archiving problems, being heavy and bulky to store (McGuinness and Stephens, 1992), which leads to a major problem in their storage. They are also liable to damage and loss, and there is difficulty in sending them to other clinicians in some cases (McGuinness and Stephens, 1992, Quimby et al., 2004). These problems have encouraged alternatives including photocopying, holography and digitised study models.

2.8.2 Construction of digital study models

The predominant method for obtaining digital models is by taking an impression. The impressions are taken at the orthodontist's office using high quality impression material and mailed to the company, which will pour the impression to create a study model that can then be scanned to create a digital model (Creed et al., 2011). Alternatively, the impression can be poured by the clinician/technician in practice and then scanned by the orthodontist if a

scanner is available in the working unit. Some examples of models created from impressions are OrthoCAD (Cadent, Inc., Carlstadt, NJ) and emodels (GeoDigm, Inc., Chanhassen, MN).

Impression materials are prone to dimensional changes due to chemical reactions and might show expansion due to secondary reactions whilst setting. This may have an impact on the dimensional accuracy of plaster study models where digital models are constructed. An intra-oral scanner could overcome some of the errors associated with traditional impression taking and cast production, as digital output data can be fed directly into a digital workflow (van der Meer et al., 2012).

Recently, advances in technology have led to digital models being created via cone beam computerised tomography (CBCT). With this technology the models are embedded in the CBCT image, which allows all anatomical landmarks captured during the scan to be viewed, e.g. roots, bone height and impacted teeth. However, this may necessitate high radiation exposure to the patient, which may need to be justified. Several research units are currently undertaking research work to overcome this side effect (Creed et al., 2011).

A systematic review published by Luu et al. (2012) reported that the available literature on 3D virtual dental study models has largely focused on those acquired by laser, while others have investigated holographic scanning, stereophotogammetry capture and, more recently, cone-beam computed tomography (CBCT). The authors investigated the impact of the different methods of constructing digital study models (laser-acquired and CBCT-acquired) for the validity and reliability of linear measurements. No perceived clinically significant differences were found in intra-examiner reliability and validity across the various acquisition types. However, the variation in

correlation for two-landmark measures from CBCT-acquired models was the only inconsistent finding. It is worth mentioning that the authors recommended further independent studies were required to confirm their findings.

2.8.3 Digital model software systems

The most commonly used software packages for measuring digital models are:

1. OrthoCAD (Cadent, Carlstadt, NY, USA)
2. emodels (GeoDigm Corp., Chanhassen, MN, USA).
3. Other systems e.g. Dig-model, Cone Probe/Digital Capillaries and Easy3D Scan.

2.8.3.1 OrthoCAD

OrthoCAD introduced digital study models in 1999. It is a patented computer model system that creates digital images of dental casts. To obtain the digital images, the images of wax-bite impressions are scanned via stereo lithography and converted into digital images that are made available for downloading by the account holder. Studies performed using OrthoCAD software to compare plaster and digital models include:

- Measurements of tooth size, tooth width, overbite and overjet (Santoro et al., 2003, Quimby et al., 2004).
- Measurements of arch length transverse dimensions, arch length, crowding, irregularity, space available and space required (Leifert et al., 2009, Goonewardene et al., 2008, Quimby et al., 2004)
- Bolton analysis, PAR Index and ABO OGS scores (Okunami et al., 2007, Costalos et al., 2005, Mayers et al., 2005, Hildebrand et al., 2008).

Three studies were designed to compare the ABO components from plaster to digital models using OrthoCAD software (Costalos et al., 2005, Okunami et al.,

2007, Hildebrand et al., 2008). These studies will be discussed in detail in the following section (Section 2.10). Furthermore, Mayers et al. (2005) was the only study conducted that measured the PAR Index scores using OrthoCAD software. The study indicated that the scores derived from digital computer based models were valid and reliable measures of malocclusion.

2.8.3.2 emodel software

emodels constructs digital model by scanning plaster model using a non-destructive laser scanning process that digitally maps the geometry of the cast's anatomy. It has been used in numerous comparisons between plaster and digital models:

- Measurements of tooth size and time taken (Horton et al., 2010).
- Measurements of PAR Index and Bolton ratio (Stevens et al., 2006, Mullen et al., 2007).

The only study that used emodel software to compare plaster and digital measurements using the PAR Index and a Bolton analysis was undertaken by Stevens et al. (2006). They indicated that there was no clinically significant difference between plaster and digital models found using emodel's software, and the results gave no indication that digital models would cause an orthodontist to make a different diagnosis of malocclusion than with plaster models. Their conclusion was that digital models are not a compromised choice for treatment planning and diagnosis (Stevens et al., 2006).

2.8.3.3 Other software programs

- Cone Probe/ Digital Capillaries was used to compare plaster and digital models in relation to tooth width, arch length and crowding (Redlich et al., 2008).
- Easy3D Scan was used to compare plaster and digital models in

linear dimensions (Keating et al., 2008).

- Cecile3 software was used to measure tooth size, overjet and overbite (Watanabe-Kanno et al., 2009).
- Dig-model (Orthoproof, Albuquerque, MN, USA) scans directly off the impression to generate digital models (Veenema et al., 2009).

2.8.3.4 Comparison between OrthoCAD and emodel software

Table 2.2 shows a general comparison between OrthoCAD and emodel software systems.

Table 2.2 Comparison between OrthoCAD and emodel software

	OrthoCAD	**emodel software**
Cost of basic file	Cheaper than emodel software	More expensive than OrthoCAD
File size	Slightly over 800 kb	About 800 kb
Technology used	Proprietary scanning process	Non-destructive scanning
Software charge	None	None
Plaster copies available for additional cost	Yes	Yes
File saved on company's web server	10 years	Indefinitely
Ships plaster to lab for appliance Fabrication	Yes	Yes
Point-to-point measurements	Yes	Yes
Curve-length measurements	Yes	Yes
Bolton analysis	Yes	Yes
Tanaka-Johnson analysis	Yes	No
Cross-sectioning tool	Yes	Yes
Visualised occlusal contacts	Yes	Yes
Virtual diagnostic setup (extra cost)	Yes	Yes

Integration with office management software	Dolphin, Vistadent, Walrus, Sirona, Practice Works Imaging, Dr. Views, Oasys, Ortho II, OrthoChart, Televox, and OrthoSesame	Dolphin, IMS, and Vistadent
Software size	8 mb	12 mb
Plaster models fabricated at a later date	Possible for a fee	Possible for a fee
Ability to create digital models from pre-existing plaster models	Yes	Yes

2.9 Research into digital model software

2.9.1 Accuracy of digital models

The potential advantages of digital models for replacing the plaster models would be lost if the validity, reliability, efficiency and ease of linear and angular measurement using digital models were not comparable to those related to plaster models (gold standard). Many studies had been conducted to compare measurements done on plaster models and digital models and found divergent results (Mayers et al., 2005, Stevens et al., 2006, Quimby et al., 2004, Bell et al., 2003, Abizadeh et al., 2012, Sousa et al., 2012)**.** These studies assessed linear and angular measurements as well as different occlusal indices including PAR, ABO OGS, ICON indices and Bolton analysis.

Sousa et al. (2012) compared linear measurements done on plaster models with a digital calliper directly on the dental casts and digitally on the digital models. Twenty sets of study models were used in this study. Fifteen anatomic dental points were identified, and a total of 11 linear measurements were taken from each cast, including arch length and width. The authors reported that linear

measurements on digital models were accurate and reproducible as there were no statistically significant differences between the measurements made directly on the dental casts and on the digital models. Quimby et al. (2004) assessed ten different linear measurements using a large sample size of 50 sets of digital and plaster models where no statistical significant difference was found. Bell et al. (2003) reported no statistically significant difference between digital and plaster models with regards to the linear measurements. In agreement, Bell et al. (2003) stated that average difference between measurements of dental casts and 3D images was 0.27 mm. This difference was within the range of operator errors (0.10-0.48 mm), which suggests that computer-based models appear to be a clinically acceptable alternative to conventional plaster models.

Moreover, Mayers et al. (2005) conducted a study to assess the validity and reliability of scoring the PAR Index on digital models using the OrthoCAD software system. The study sample consisted of 48 pairs of plaster and digital pre-treatment models. One examiner, calibrated in the PAR Index, scored the digital and plaster models. The overall PAR scores were examined for reliability and validity by using analysis of variance and the intraclass correlation coefficient (ICC). The authors reported that PAR scores derived from digital models were valid and reliable measures of occlusion. In agreement, Stevens et al. (2006) using 25 set of plaster and digital models with different types of malocclusion using emodel software system reported that the PAR analysis and its constituent measurements were not significantly different clinically when compared to plaster models.

On the contrary, Abizadeh et al. (2012) reported statistically significant differences for occlusal features and measurements between digital and plaster models. A relatively large sample of one hundred and twelve sets of study

models with a range of malocclusions and various degrees of crowding were used. The authors stated that, in eight of 16 occlusal features measured, the plaster measurements were more repeatable, indicating that the digital model scans were not a true representation of the plaster models. In agreement, Leifert et al. (2009) in a similar study design reported a statistically significant difference between maxillary plaster and digital models when assessing space analysis.

Moreover, Tomassetti et al. (2001) studied the accuracy and efficiency of measuring Bolton's tooth-size analysis using manual measurements with a vernier calliper and three computerised methods, including OrthoCAD. Although the authors found no statistically significant differences among the tested methods, there were clinically significant differences (1.5 mm) for all methods. They concluded that OrthoCAD was among the methods with the greatest differences. In contrast Santoro et al. (2003) reported a statistically significant difference when measuring tooth size, overbite and overjet using OrthoCAD models compared with plaster models; however, the authors suggested that the differences were considered to be clinically insignificant (0.5 mm).

The divergent results reported from the above-mentioned studies indicate the need for a systematic review of the literature in order to evaluate the available evidence. Fleming et al. (2011) undertook a well-designed systematic review to evaluate the validity of the digital models by assessing measurements done on both digital and plaster models. The systematic review was properly designed using the PRISMA guidelines with a structured search strategy. The authors included studies that assessed both linear and angular measurements as well as different orthodontic indices e.g. PAR, ICON and ABO OGS. Seventeen

articles out of 283 were included in the review; however, meta-analysis was not conducted due to the heterogeneity of the included studies. It was reported that, overall, the absolute mean differences between direct and indirect measurements on plaster and digital models were minor and clinically insignificant and that orthodontic measurements with digital models were comparable to those derived from plaster models with a higher degree of validity when compared with plaster models. However, these results should be interpreted with caution as the evidence identified in this review is of variable quality, as reported by the authors. It is also important to highlight that the authors (Fleming et al., 2011) instigated the assessments of validity but not the reliability of the measurements on digital models.

In 2012, another adequately designed systematic review was published which aimed at assessing the validity and reliability of linear measurements only on digital models (Luu et al., 2012). A structured search strategy following the PICO model was conducted with strict inclusion criteria. Only 17 studies out of 278 were included in the review, with three reviewers assessing the studies independently. No meta-analysis was conducted due to the heterogeneity of the included studies .In agreement with Fleming et al. (2011), Luu et al. (2012) reported that the validity of the linear measurements of digital models was clinically acceptable compared with plaster models. In addition, the authors reported clinically acceptable intra-rater reliability for both the digital and plaster study models with regard to linear measurements.

Well-designed systematic reviews are considered to be at the highest level of scientific evidence. According to the reported results from the above-mentioned systematic reviews (Luu et al., 2012, Fleming et al., 2011) there seems to be some evidence to suggest that measurements conducted on digital models are

considered to be reliable and accurate enough for clinical use when compared to plaster models.

2.10 Studies comparing plaster and digital models measuring ABO OGS components

Three studies have been conducted to compare the ABO OGS components from plaster and digital models using OrthoCAD software (Costalos et al., 2005, Okunami et al., 2007, Hildebrand et al., 2008).The purpose of these three studies was to determine whether the ABO OGS can be reliably, consistently and accurately assessed from digital casts, and whether there are statistically significant differences between digital and plaster dental casts in ABO OGS scores.

In the ABO OGS, eight occlusal criteria are recorded: tooth alignment, vertical positioning of marginal ridges, buccolingual inclination of posterior teeth, occlusal relationship, occlusal contacts, overjet, interproximal contacts and root angulation. The first seven can be assessed on orthodontic study models, which can be either plaster or digital. The eighth criterion, root angulation, is measured on a panoramic radiograph.

When using the plaster models, a special instrument called the ABO measuring gauge has been developed for assisting in taking these measurements (Figure 4.1). On the other hand, the technique used to measure digital models aims to utilise the software used in each study performed. After measuring the seven components of the ABO OGS separately, each score is then used to compare between plaster and digital models. Then, from the scores for each of the criteria (excluding the eighth criterion), a total score is given that is the sum of all criteria, and this is then used as another comparison between plaster and digital models.

2.10.1 Sample size used in studies conducted on ABO OGS software systems

Costalos et al. (2005) evaluated the accuracy of digital model analysis for the ABO OGS using 24 models which were taken from patients at the completion of their orthodontic treatment at the postgraduate orthodontic clinic of the Columbia University School of Dental and Oral Surgery and sent to OrthoCAD to construct plaster and digital models. These models were selected according to only three criteria:

- No deciduous teeth were present.
- No edentulous spaces were present.
- Acceptable molar and canine relationships, overjet and overbite on visual inspection.

Okunami et al. (2007) assessed the accuracy of ABO OGS scoring using digital models when compared to plaster models using 30 post-treatment plaster casts which were selected from the University of Illinois at Chicago's Department of Orthodontics. These models were selected according to four criteria:

- Patients who had nonsurgical comprehensive orthodontic treatment with fixed appliances.
- The dental casts had to be in an acceptable condition with all incisors, canines, at least one premolar, and the first and second molars bilaterally in both the maxillary and mandibular arches.
- No duplicates of the dental models.
- No consideration of age and sex.

Finally, Hildebrand et al. (2008) undertook a study to evaluate the accuracy of digital ABO OGS scoring compared to scoring on plaster models. They used a slightly larger sample consisting of 36 randomly selected plaster models from

finished orthodontics cases. These models were selected according to two criteria:

- Good condition casts without bubbles or broken teeth.
- Properly trimmed backs flush in a way that when the casts were placed on flat surface, they showed maximum interception occlusion.

It is obvious from the above-mentioned studies that a sample size ranging between 24-36 study models were used for comparing ABO OGS scoring using the digital and plaster models (Table 2.3). It is important to mention that none of these studies conducted a prior sample size calculation to decide on the appropriate sample required for the statistical analysis conducted.

2.10.2 Methods used in studies conducted on ABO OGS software systems

In the Costalos et al. (2005) study, two examiners were involved in the assessment. Examiner 1 was a postgraduate student and was trained with the voice-over CD-ROM provided by the ABO and information about the ABO grading available in the OrthoCAD version 2.17 software package. The seven criteria of the ABO objective grading system were scored on the 24 models, and four weeks later the plaster models were scored. Subsequently, a second analysis was done for each patient by using the digital models grading software and the same criteria and grading tool. The scores of the seven criteria were summed up to drive a total score per model.

A second examiner, a senior pre-doctoral dental student at Columbia, was trained in the same way as Examiner 1. Calibration of the two examiners was done by having each score result from a selection of two plaster models and two digital model OrthoCAD models independently, and then jointly review their scorings after each model analysis. Examiner 2 repeated the analyses on

the 24 plaster models and 24 digital models inspected by Examiner 1.

Similarly, the Okunami et al. (2007) study involved two examiners. The primary examiner (background not mentioned) was trained by a former ABO examiner who was familiar with the models analysis. A training session was conducted to teach the examiner to score plaster models and establish consistency measurements. After the initial training, inter-examiner calibration was conducted between the examiner and the former ABO examiner. Five sets of plaster models were used, and measurements were made separately by each person and compared. If there was a difference of more than two points per component, measurements were repeated again.

There was no official session with the ABO program of OrthoCAD. The primary investigator followed OrthoCAD's instructions. An intra-examiner calibration was also conducted to establish the consistency of the measurements. Ten plaster models and their corresponding digital models were selected at random. Two plaster and digital models were measured per day for five consecutive days. The same procedure was repeated two weeks later, and measurements were compared. After calibration, all 30 plaster and digital models were measured. Measurements were made on two randomly selected plaster models each day until all models had been measured.

Unlike the two previously mentioned studies in this section, Hildebrand et al. (2008) used only one examiner (background not mentioned). Intra-examiner reliability was determined by scoring ten randomly selected models three times, two weeks apart. Training of the examiner in the use of the ABO gauge was performed with the ABO calibration kit, the voice-over CD-ROM from ABO, and the information for ABO available in the OrthoCAD software package.

It is notable that in both Costalos et al. (2005) and Hildebrand et al. (2008)

studies the examiners received training to use both the digital software scoring and the plaster model scoring. However, in Okunami et al.'s (2007) study training was only provided to the examiners on the plaster models with no official training for the digital scoring using the OrthoCAD. This may have an influence on the scoring reliability of the examiners.

It is worth mentioning that both Costalos et al. (2005) and Okunami et al. (2007) undertook the assessment of intra-examiner and inter-examiner reliability, as both studies had two examiners scoring the models, while Hildebrand et al. (2008) had only one examiner, so only intra-examiner reliability was reported.

All three studies mentioned above used the OrthoCAD software system for scoring the ABO OGS on digital models. However, different versions of the software were used in Costalos et al. (2005) and Hildebrand et al. (2008) studies (version 2.17 and 2.66, respectively), while Okunami et al. (2007) did not specify the version code for the software used (Table 2.3) . It is not clear if there were significant differences between the different versions of the OrthoCAD software that could have an impact on the results of each study.

2.10.3 Data analysis in studies conducted on ABO OGS software systems

Costalos et al. (2005) applied a statistical analysis to investigate the accuracy and repeatability of the methods studied with the aid of the statistical program SAS for Windows Version 8.0. Means and standard deviations of the seven variables and the total score for plaster and digital models were calculated separately. Means and standard deviations of the difference of the total scores between plaster and digital models were also calculated. Analysis of variance (ANOVA) was applied to the data to determine whether the two methods had

equivalent means for the seven variables and the total score. An interclass correlation coefficient of reliability was calculated for each examiner in order to assess the reliability of digital models compared with plaster models. Inter-examiner error was evaluated, and the mean values of the measurements by each examiner were compared. Fixed and random examiner models were used.

Okunami et al. (2007) entered data into a data file for statistical analysis by using the Statistical Package for the Social Sciences (SPSS). A detailed analysis was conducted for investigator reliability and the data collection of 30 plaster and digital models. For investigator reliability, both plaster and digital models were measured on two separate occasions. A comparison was done separately for plaster models and digital models, and the Wilcoxon rank-test was used to evaluate statistical differences. The results suggest the consistency of the measurements for each criterion and the total score of ABO OGS components by the investigator. In addition, a detailed analysis of the data was conducted by using the Wilcoxon rank-test to determine any statistically significant differences between plaster and digital dental models.

Hildebrand et al. (2008) used data from 36 digital and plaster casts entered into an Excel sheet and subsequently transferred the information to SPSS software. The scores for each of the 7 components of the 36 casts were individually compared, as were the total scores. Ranges, absolute means and standard deviations were calculated for all measurements taken. Calculating the absolute difference of each model allowed the examiner to see the total variation that was actually present. The Spearman rank correlation coefficient was also computed in order to assess intra-examiner reliability. Ordinal indices have been analysed by using parametric tests instead of nonparametric tests. Three types of statistics were run to compare the digital casts with the plaster model

casts.

First, a descriptive analysis was reported for the absolute differences between the plaster and digital casts for each subject in each component and the total ABO OGS score. Descriptive statistics of the differences included range, absolute mean and standard deviation. Second, the Spearman rank correlation coefficient was calculated for each component and the total ABO score in order to assess the degree of association. Third, the Wilcoxon rank-sum test was used to find differences in the 7 ABO OGS scoring components.

Table 2.3 Studies that investigated different components of the ABO score

Year	**Authors**	**Primary Outcome**	**Method**	**Sample Size**	**Software**	**Statistical Test**
2005	Costalos et al.	ABO OGS components and total scores	Plaster SM vs. Digital SM	24 Post-treatment models	OrthoCAD Version 2.17 software package	• Correlation coefficient • ANOVA test
2007	Okunami et al.	ABO OGS components and total scores	Plaster SM vs. Digital SM	30 Post-treatment models	OrthoCAD's ABO OGS Software program	• Wilcoxon rank-sum test
2008	Hildebrand et al.	ABO OGS components and total scores	Plaster SM vs. Digital SM	36 Post-treatment models	OrthoCAD's ABO OGS Software program version 2.66	• Spearman rank correlation coefficient • Interclass correlation coefficient • Wilcoxon rank-sum test • Paired Wilcoxon rank-sum test

2.10.4. Reliability of examiners in studies conducted on ABO OGS software systems

Costalos et al. (2005) used an ANOVA test to examine whether there were differences between the scores of the two examiners involved in the study. The *P* value was less than 0.0001, which demonstrates that there were statistically significant differences between the two examiners. However, the reliability was slightly higher for digital models than plaster models, R = 69 and R = 0.53 respectively. Hildebrand et al. (2008) only used a single examiner in the study, which did not allow for an assessment of inter-examiner reliability of the two methods. AlthoughOkunami et al. (2007) involved two examiners in the study the authors did not conduct a statistical test to investigate the inter-examiner reliability for the digital or plaster models' scoring methods.

While Costalos et al. (2005) reported no data related to intra-examiner reliability, Okunami et al. (2007) and Hildebrand et al. (2008) reported high intra-examiner reliability for both the digital and plaster models.

It can be concluded from the above-mentioned studies that the intra-examiner reliability of both plaster and digital models is high, while there is little evidence to suggest that inter-examiner reliability is low.

2.10.5 Overall evaluation of the results in studies conducted on ABO OGS software systems

The study of Costalos et al. (2005) showed that the plaster and digital models analysed by Examiner 1 had a total score mean difference of 1.5points, which was found to be statistically insignificant (*P*=0.3467). There was a high correlation between the total scores for both models. This finding was supported by a high interclass correlation coefficient of reliability for the total score for both examiners (Examiner 1: R=0.69; Examiner 2: R=0.86).

In addition to the total ABO OGS score, the mean differences of five components of the ABO OGS (marginal ridges, occlusal contacts, occlusal relationship, overjet and interproximal contacts) between the scores of the plaster models and digital models were found to be statistically insignificant ($P>0.05$). However, the means for the alignment component were significantly different between plaster models and digital models ($P<0.05$). In addition, the *P* value for buccolingual inclination was marginally significant with $P=0.0507$. Therefore, the authors reported that both plaster and digital models had a good correlation, except for alignment and buccolingual inclination.

Costalos et al. (2005) stated in their study that the significant difference found in the alignment and buccolingual components might be due to difficulty in identifying the same landmarks on plaster and OrthoCAD models. The limitations in significantly enlarging digital models might have contributed to this problem. This in agreement with Horton et al. (2010),who reported that one of the greatest sources of random error is the difficulty in identifying landmarks.

During the analysis of Costalos et al. (2005), two concerns were noted. Generally, a macroscopic assessment of alignment was performed, and the amount of misalignment of adjacent teeth was measured with an ABO measuring gauge. However, the same measurements performed on the digital models gains in microscopic detail. Two points were placed on each tooth to assess alignment. Slightly moving each point does not significantly change the visual microscopic alignment of the adjacent teeth. However, points might be deducted or spared. There appears to be a range where points can be spared or deducted depending on the ultimate position of the two selected points, even though the resulting line of alignment is in an acceptable position. Therefore,

Costalos et al. (2005) stated that the examiners might have had difficulty in pinpointing the exact mesial and distal points that could be used to evaluate alignment, making it difficult to consistently and accurately make measurements.

When measuring buccolingual inclination, Costalos et al. (2005) noted a variation with OrthoCAD. The buccolingual inclination of the tooth is assessed from a plane created from a cusp tip on that tooth extended to the cusp tip of a collateral tooth. Ordinarily, contra lateral human teeth are not positioned parallel to each other. This is particularly true of premolars, which converge toward the anterior midline. Therefore, Costalos et al. (2005) stated that it is difficult to assess buccolingual inclination on digital models with only a single line because this line cannot symmetrically bisect the occlusal surface of each contra lateral tooth at the cusp tip. They also stated that measurements on plaster models are different because the plaster models are truly three-dimensional; therefore, buccolingual inclination is visualised from the plane created by the ABO measuring gauge. Accordingly, they stated that this variation can be avoided if this plane is formed by two lines joined to form an angle at the sagittal plane. If these lines are maintained in the same plane, they can be manipulated by the examiner to bisect each contra lateral tooth at the cusp tip and therefore more accurately measure buccolingual inclination.

The total ABO OGS score of Okunami et al. (2007) was statistically significantly different for digital models compared with plaster models (P=0.000). Occlusal contacts and occlusal relationships were also statistically significantly different (0.000 and 0.023, respectively). However, the mean difference for the alignment, marginal ridges, overjet and interproximal contacts components were found to be statistically insignificant. The authors

concluded that the OrthoCAD software was not adequate for ABO OGS scoring.

Okunami et al. (2007) stated that a major problem was encountered throughout the study that explained the difference for occlusal contacts and total score. When the examiners performed the occlusal contacts measurements, the digital images of the maxillary and mandibular teeth overlapped each other. The result was an occlusal contact measurement that measured the amount of vertical overlap between images, rather than the distance that the teeth were not in contact. Therefore, points were deducted unnecessarily.

The authors stated that OrthoCAD was aware of the problem and tried to fix it by manipulating the image. However, instead of solving the problem, another problem was encountered, which was that several models no longer had occlusal contacts. When they were made aware of this problem, OrthoCAD conceded that the problem could not be fixed with the software version they had used for their study. Because the number of points deducted for occlusal contacts was so substantial, the total scores calculated for the digital models were also greatly affected.

In addition, the authors explained that the statistically significant difference found between the two scoring methods in the occlusal relationship component was due to discrepancy while measurements were taken. If the plaster models were not viewed perpendicularly, the measurements might have been interpreted differently. In contrast, the digital model could be viewed from the exact angulation by using the predefined spots on the horizontal plane in the view.

Hildebrand et al. (2008) reported that the total score of the ABO OGS measured digitally was found to be statistically significantly different from the

plaster model measurements (*P*=0.001). The digital scoring was on average 9 points greater, with a range of -1 to 21 points, which may also suggest a clinically significant difference. Among the 7 components that comprise the total score, 3 (alignment, occlusal contacts, overjet) showed statistically significant differences. The greatest clinically significant difference was in overjet, with an average difference of 3.94 points. The authors explained that the 9-point average increase difference in the digital scoring was due to statistically significant differences in alignment, overjet and occlusal contacts that appeared to be related to 2 factors. The first factor was a combination of systematic errors in the software algorithm used to compute the scores, and the second was the difference in the computerised articulation of the digital casts.

In summary, the results from both Hildebrand et al. (2008)and Okunami et al. (2007) of the digital scoring were found to be statistically significantly different from the plaster model measurements for ABO OGS scoring. In contrast, Costalos et al. (2005) reported no statistically significant difference between the two methods in the majority of the ABO OGS components (Table 2.4).

It was noted that the above-mentioned studies did not agree on the components of the ABO OGS that caused the difference between the plaster and digital scoring, which may be explained by the difference in sample size, statistical tests and versions of the OrthoCAD software among the three studies. Therefore, it can be concluded that OrthoCAD software is not an accurate method for measuring the ABO OGS. This may suggest that further studies needs to be conducted using other ABO OGS software systems.

Table 2.4 Studies that investigated different components of the ABO score

ABO OGS components	Costalos et al. (2005)	Okunami et al. (2007)	Hildebrand et al. (2008)
Alignment	Statistically significant difference <0.0001	No statistically significant difference 0.340	Statistically significant difference 0.001
Marginal Ridges	No statistically significant difference 0.4694	No statistically significant difference 0.837	No statistically significant difference 0.107
Buccolingual Inclination	Statistically significant difference 0.0507	Not included in the study	No statistically significant difference 0.564
Occlusal Relationship	No statistically significant difference 0.3567	Statistically significant difference 0.023	No statistically significant difference 0.414
Occlusal Contacts	No statistically significant difference 0.2169	Statistically significant difference 0.000	Statistically significant difference 0.032
Overjet	No statistically significant difference 0.1077	No statistically significant difference 0.100	Statistically significant difference 0.001
Interproximal Contacts	No statistically significant difference	No statistically significant difference	No statistically significant difference

	0.0613	0.102	0.317
Total Score	No statistically significant difference 0.3467	Statistically significant difference 0.000	Statistically significant difference <0.001

2.11 Time taken to compare plaster and digital models

The time factor is an important element to be considered when evaluating the clinical effectiveness of a system. When the digital models were introduced into the dental field it was assumed that they could overcome several limitations of the plaster models including saving the clinicians valuable time, which can justify the cost of implementing the system. Several studies have been conducted to compare the time taken for different measurements and indices scoring between plaster and digital models (Mayers et al., 2005, Mullen et al., 2007, Horton et al., 2010).

Mayers et al. (2005) assumed that, when comparing plaster and digital models for peer assessment rating PAR scoring using the OrthoCAD CRT software (Cadent, Carlstadt, NJ), scoring digital models would be faster than scoring plaster models. However, the authors took about 44 minutes to score 10 sets of plaster models, but required 70 minutes to score 10 sets of digital models. It is important to note that the examiner was PAR calibrated on the plaster models but did not receive proper training in the use of the new software. This may explain the prolonged time taken to score the PAR Index using the new software. Interestingly, the authors stated that measurement times did not progressively decrease due to any learning effect; they indicated that the difference in measurement times was because the software was not amenable to efficiently scoring a model. Overjet of the four incisors was the most time-

consuming component in their study. However, another factor may influence the relative time required to compare the two types of models: the time recorded did not include time spent retrieving and refilling the plaster models, whereas the digital models were immediately available.

Mullen et al. (2007)compared emodels (version 6.0, GeoDigm Corp., Chanhassen, Minn) to plaster models in measuring the accuracy of the Bolton ratio and the time to perform a Bolton analysis for each point, which was recorded in seconds. The emodels software was faster than the plaster models by an average of 1 minute and 6 seconds. This difference was found to be neither statistically nor clinically significant. The authors suggested that the reason emodels was faster than plaster models was because, when measuring the Bolton ratio on plaster models, they had to record results on paper and then calculate the measurements with a calculator, in contrast to the emodels calculations, which were performed at a click of a button. This afforded the emodels an edge related to time involved in calculating the Bolton ratio.

Another study was designed to compare the time taken to compare the plaster and digital models; this study involved the measurement of mesiodistal tooth width(Horton et al., 2010) using emodels (GeoDigm, Chanhassaen, Minn) software. The authors used five different techniques: occlusal aspect, occlusal aspect zooming in each individual tooth, facial aspect rotating as needed, facial aspect from three standard positions, and qualitatively rotating the model in any position deemed necessary. Each set of measurements was timed to the nearest second using a stopwatch. The average time taken to measure plaster casts was 4 minutes 15 seconds, and the average time taken to measure digital models was 2 minutes for occlusal aspect, 2 minutes 29 seconds for occlusal aspect zooming, while the facial aspect rotation method took 4 minutes 21 seconds,

the aspect from three standard positions method took 1 minute 51 seconds, and the qualitative technique took 7 minutes. The authors indicated that both qualitative and occlusal techniques had strong repeatability and accuracy, and both were acceptable measurement techniques. However, the authors recommended using the occlusal technique, as less time was required compared with the plaster models (plaster models = 4 minutes 15 seconds vs. occlusal technique = 2 minutes).

It is obvious that there are a limited number of studies reported in the literature that investigated the time taken when using the digital models for linear measurements or scoring orthodontic indices. It seems that for simple linear measurements digital measurements can be faster. However, on the level of using sophisticated occlusal and outcome indices there seems to be no significant advantage of using digital models in regards to the time taken for examiner to score the index.

2.12 The new software system

The new software named ABO is a three-dimensional piece of software developed in collaboration between Bioprecision diagnostics (http://www.bioprecision.co.uk/) and the University of Dundee. It has nothing to do with the American Board of Orthodontics. The software aims to measure the total score and components of ABO OGS except the root angulation component, which is measured on the x-rays, to replace the conventional method of measuring the ABO OGS components on plaster models.

After the process of scanning the plaster models are inserted into the ABO software by scanner supplied by 3shape scanners (http://www.3shape.com/) the models are saved by name and a user can select which model they want to

measure. The ABO software has several advantages that can facilitate the measuring process:

1. Zoom
2. Pan
3. Rotate
4. Indicate point

The digital models in the software can be viewed as a full arch or upper and lower arches alone, while a cross section can also enable transverse, occlusal and centerline views .The software measures all components of the ABO OGS separately by indicating points in the software guide to the user about which point should be indicated. After measuring the 7 components of ABO OGS the total score is added automatically and the data is then saved on the software and the user can move to the following model to measure.

2.13 Summary of the literature review

It seems that there is an obvious trend towards the use of digital models in association with different software systems to electronically score different orthodontic treatment outcome indices including the ABO OGS. Some articles have been published to investigate the validity and reliability of using different versions of OrthoCAD software when compared to the plaster model conventional method (gold standard). It can be concluded that OrthoCAD software is not an accurate method for measuring the ABO OGS with clinically and statistically significant differences when compared to the conventional method .The intra-examiner reliability of digital models using OrthoCAD was found to be high, while there is little evidence to suggest that inter-examiner reliability is low. Moreover, it was reported that there is no advantage in using

the OrthoCAD software to reduce the time for scoring the ABO OGS. It has been suggested that technical problems with the OrthoCAD software systems might have been the main reason for the reported discrepancy. New software systems could be developed with an attempt made to provide the clinician with accurate and reliable scoring of the ABO OGS.

Chapter 3: Aim and Hypotheses

3.1 The aim of the study

Primary aim:

To investigate the validity and reliability of a newly developed piece of software in order to score the different components of the American Board of Orthodontics Objective Grading System (ABO OGS) using digital models to replace the conventional method (gold standard) of using plaster models and a hand ABO gauge to score the ABO OGS.

Secondary aims:

- To assess the level of agreement between orthodontists and non-orthodontists' ABO OGS scores using the conventional method on plaster models and the new software on digital models.
- To compare the time taken by examiners to score the ABO OGS using the conventional method on plaster models and the new software on digital models.

3.2 Null hypotheses

The null hypotheses to be tested were:

Ho1: There is no agreement with the American Board of Orthodontics Objective Grading System (ABO OGS) scores (total score and individual components) between measurements that recorded using the new software on digital models and used the conventional method on plaster models.

Ho2: There is no difference in the intra-examiner and inter-examiner reliability of ABO OGS scores (total score and individual components) between the measurements recorded using the new software on digital models and the conventional method on plaster models.

Ho3: There is no difference in the ABO OGS scores (overall and individual components) between specialist orthodontists and non-orthodontists.

Ho4: There is no difference in the time taken to measure the ABO OGS between ABO software on digital models and the conventional method on plaster models.

Chapter 4: Materials and Methods

4.1 Materials

Thirty-one post-treatment scanned study models were used in this study. Fifteen models which met the inclusion criteria were randomly selected from a private practice and sixteen models were selected from the University of Dundee Dental Hospital. All cases had completed fixed-appliance orthodontic treatment in both mandibular and maxillary arches.

4.1.1 Inclusion criteria for the study models

The inclusion criteria for the study models used in this investigation included:

1. Completed fixed-appliance orthodontic treatment in both mandibular and maxillary arches.
2. All incisors, canines and at least one premolar present per quadrant
3. No broken dental casts.
4. Properly trimmed casts showing correct occlusion.
5. The models were selected with no consideration of age or gender.

4.1.2 Scanning and digitisation of the models

After the selection of the thirty one plaster models, the models were then sent to the Bioprecision Diagnostics Company (http://www.bioprecision.co.uk) for scanning using a digital laser scanner (3shape Scan IT Orthodontics) and were then uploaded on the new software. The following steps were performed by the new software company:

1. A high-resolution scan was made of each arch (upper and lower) individually. For these scans, the arches were mounted on a base plate that fitted into the scanner.
2. A low-resolution scan was made of the arches mounted in occlusion using a clamp-like mount that fitted into the scanner.
3. The scanning software presented the operator with high-resolution scan

arches and the low-resolution scan occlusion. The operator indicated common points on the different scans to instruct the software how to align the high resolution scans properly in occlusion.

4. The high-resolution scans were then saved. The low-resolution scan was discarded. The operator then loaded the digital models into another piece of software called Rhinoceros (http://www.rhino3d.com/). This enabled the operator to tidy (trim) parts of the models that were not wanted, such as any aspects of the mount used to hold the plaster model in the scanner that might be present. The trimmed digital models were then saved.
5. A side-by-side comparison of the digital and plaster models was performed, normally by someone other than the operator who had scanned and trimmed the digital models. This was to ensure that the digital models fairly represented the physical models in their occlusion and alignment, and that the trimming process had not removed any part of the models that should be retained.
6. If the digital models were not sufficiently accurate, then depending on whether it was the alignment, the trimming, or even simply the quality of the scan that was at fault, the process reverted to an appropriate point and was repeated from that point.

After the scanning process was done the scanned models were uploaded on the new software and each set of models was coded containing the upper and lower arch.

Three plaster models provided by ABO for calibration were used for training and calibration purposes using the ABO gauge (Figure 4.1).

4.1.3 The ABO Gauge (Casko et al., 1998)

(A) This portion is 1 mm in width and used to measure discrepancies in alignment, overjet, occlusal contact, interproximal and occlusal relationships.

(B) This portion has steps measuring 1 mm in height and is used to determine discrepancies in mandibular posterior buccolingual inclination.

(C) This portion has steps measuring 1 mm in height and is used to determine discrepancies in marginal ridges.

(D) This portion has steps measuring 1 mm in height and is used to determine discrepancies in maxillary posterior buccolingual inclination.

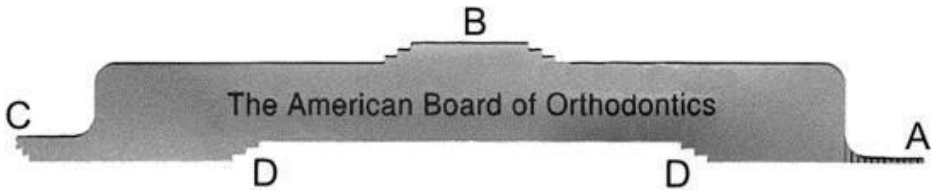


Figure 4.1 ABO gauge (Casko et al., 1998)

4.1.4 Examiners involved in the study

The four examiners were selected to have different levels of orthodontic knowledge and experience to investigate if non-orthodontists can yield reliable ABO OGS scores using plaster and digital models when compared to orthodontists.

- Examiner 1: Orthodontic postgraduate research student who was the most trained in the current study on the ABO OGS and had undertaken research related to area of the occlusal indices used in orthodontics.
- Examiner 2: Orthodontist had more than 8 years of experience in the field and was familiar with the ABO OGS index.

- Examiner 3: Prosthodontics postgraduate student (second year) who was not familiar with the ABO OGS index before participating in the current study.
- Examiner 4: Undergraduate dental student (last year) who was not familiar with the ABO OGS index before participating in the current study.

4.1.5 The new software system

New software was developed in collaboration between Bioprecision Diagnostics and the University of Dundee to measure ABO OGS on digital study models in order to meet the need for valid and reliable ABO OGS measurement software. The software is not commercially available but was made available to investigators in this study so as to assess its validity and reliability.

4.2 Methods

4.2.1 Examiners' training and calibration

Plaster model training

The study examiners were trained on measuring plaster models by watching videos on the ABO website, where a former ABO examiner demonstrated how to score the ABO OGS using plaster models (http://www.americanboardortho.com/professionals/clinicalexam/casereportpresentation/preparation/measurement_demo.aspx). An ABO calibration kit was ordered from The American Board of Orthodontics in the United States of America for training purposes. This included three plaster models, which were sent with the ABO gauge (Figure 4.1), and scoring sheets. The examiners were trained on the plaster models by using the ABO calibration kit. In addition Examiners 2, 3 and 4 were trained with the help of Examiner 1, who had the

most experience in scoring ABO OGS using plaster models.

After watching the training video, all examiners scored each calibration model for each component of the ABO OGS. The examiners did not have access to the results that were given in the calibration kit. The sheet developed by the ABO was used for recording measurements. Training continued on the calibration models for as many repetitions as the examiner required until each of them had the same scores as located on the result sheet in the calibration kit.

The new software training

The new software was used for scoring the digital models. Examiner 1 was trained to measure the digital models by following the instructions of the new software package. After repetitive use of the new software, Examiner 1 helped the rest of the examiners to be familiar with the use of the new software. No calibration was performed for the examiners, as the manufacturers did not supply a calibration kit. The software indicated the point that should be taken automatically on each tooth digitally in order to measure each component. Then, the software automatically calculated each component and the sum of the seven components of the ABO OGS to give a total score.

4.2.2 Examiners' scoring

All examiners scored the 7 components of the ABO OGS (alignment, marginal ridge, buccolingual inclination, occlusal contacts, occlusal relationship, overjet, interproximal contacts) for the 31 plaster models involved in the study once on these models. Examiner 1 scored the 31 models twice with an interval of two weeks to assess the intra-examiner reliability. The plaster models were scored using the ABO hand-measuring gauge (Figure 4.1), and the measurements were plotted on the ABO OGS evaluation sheet.

4.2.3 Software advantages

The ABO software had the ability to show the upper or lower arches and both arches together on the screen. Cross sections could be enabled to show transverse, occlusal and centreline views. ‘Zoom Pan’, ‘Rotate’ and ‘Indicate Point’ were options that were found on the software that could facilitate measuring the models digitally (Figures 4.2 and 4.3).

Figure 4.2 Screenshot of a digital model using the new software

Figure 4.3 Rotating the digital model using the new software

4.3 Statistical analysis

Data was then uploaded into an Excel spreadsheet and imported to the Statistical Package for the Social Sciences (SPSS) (IBM SPSS Statistics 19)) for statistical analysis.

4.3.1 Descriptive analysis

For each outcome variable of the two scoring methods, descriptive statistics were performed. For continuous variables, the means and standard deviations were calculated for each group.

4.3.2 Examiner reliability

The correlations between examiners were assessed using the Pearson correlation coefficient for all components of the ABO OGS, as well as the total score for both plaster and digital models. Pearson correlation is a measure of the strength and direction of the linear relationship between two variables, where the calculated R value ranges from -1 to +1.

Intra-examiner reliability: Data from the two episodes of Examiner 1 scores was used to assess the intra-examiner reliability for both the plaster and digital models.

Inter-examiner reliability: Data from the four examiners scores was used to assess the inter-examiner reliability for both the plaster and digital models.

4.3.3 Agreement between methods

Bland-Altman plots were used to assess the level of agreement between ABO OGS scores from both plaster and digital models. To construct the Bland and Altman plots it was necessary to calculate the mean score and the mean difference between the plaster and digital models for each examiner. The limits of agreement were calculated from the standard deviation and multiplied by 1.96.

The mean of scores of the repeated measurements of Examiner 1 was used for constructing the Bland and Altman plot for assessing the agreement between plaster and digital models. While for Examiners 2, 3, and 4 a single score for each examiner was used as they only measured the plaster and digital models once.

4.4 Time

4.4.1 Methods

A total of 124 records were collected for each of the plaster and digital models (31 models assessed by four examiners) to calculate the time taken in minutes to score the ABO OGS for each set of models. The data was loaded into Excel and subsequently loaded into SPSS (IBM SPSS Statistics 19) for statistical analysis.

4.4.2 Time statistical analysis between the two methods

- A Shapiro-Wilk test was used to test the normality of the data.
- The t-test was used to analyse the difference in time between the two methods for measuring the ABO OGS total score.
- ANOVA was used to analyse the differences in time taken for scoring among the four examiners involved in the study.

Chapter 5: Results

5.1 Descriptive analysis

5.1.1 Mean, standard deviation and range for plaster models

Table 5.1 shows the results of the descriptive analysis of the scores for the four examiners for each component of ABO OGS on the plaster study models. The ABO OGS total score mean showed a range of 21.38 to 22.143 for the four examiners. Examiner 1 showed the highest mean (22.143) for the total score while Examiner 4 showed the lowest mean (21.387). Interproximal contact component showed the lowest mean points deducted compared to the other ABO OGS components and this result was indicated by the four examiners for plaster models. The lowest points mean deducted was for Examiner 4 and the highest was Examiner 1.

The buccolingual inclination component showed the highest points deducted compared to the other ABO OGS components and was indicated by the four examiners (Table 5.1). Examiners 3 and 4 showed the same mean of points deduction and had a higher mean than Examiners 1 and 2 for the buccolingual inclination component. The occlusal contact component for all examiners showed a mean deduction of less than 3 points. Examiners 2 and 3 showed the same mean for occlusal contact deduction.

The alignment, marginal ridge, occlusal relationship and overjet for the four examiners showed a mean range of between 3 to 4 points deduction. Examiner 1 showed the highest points deducted in the alignment component with a mean value of 3.532; on the other hand, Examiner 4 showed the lowest points deducted with a mean value of 3.290.

Examiner 2 showed the highest points deducted in the marginal ridge component with a mean value of 3.161; on the other hand, Examiner 4 showed the lowest points deducted with a mean value of 3.065. Examiner 2 showed the highest points deducted in the occlusal relationship component with a mean value of 3.774; on the other hand,

Examiner 4 showed the lowest points deducted with a mean value of 3.548. Examiner 4 showed the highest points deducted in the overjet component with a mean value of 3.774; on the other hand, Examiner 3 showed the lowest points deducted with a mean value of 3.323.

5.1.2 Mean, standard deviation and range for digital models

Table 5.2 shows the results of the descriptive analysis of the scores for the four examiners for each component of ABO OGS on the digital study models. The total score of the points deducted by digital models when measuring the ABO OGS showed a high points deduction with a range of (66 to 148). Examiner 4 showed the lowest points deduction with a range of (66 to 133) and a mean of 89.323. On the other hand, Examiner 2 showed the highest points deduction for the total score with a range of (90 to 148) with a mean of 124.097.

The buccolingual inclination showed a range of points deducted for the four examiners of (0 to 19). Examiners 3 and 4 showed the lowest mean points deducted, while Examiner 1 showed the highest mean points deducted among the four examiners. The interproximal contact ABO OGS component and overjet showed the highest points deducted for digital models. The interproximal contact component showed a range of points deducted for the four examiners of (2 to 41), where Examiner 4 showed the lowest points deduction with a mean of 6.548. On the other hand, Examiners 2 and 3 showed the highest points deduction for interproximal contact.

The overjet component showed a high mean deduction for the four examiners with a range (11 to 34) points. Examiner 1 showed the lowest points deduction with a range of (15 to 33) with a mean of 24.080.

The alignment component showed a range of points deducted for the four examiners of (1-31), where Examiner 2 showed the lowest points deduction with a range of (2-

19) with a mean of 7.677. On the other hand, Examiner 4 showed the highest points deduction for alignment component with a range of (1-29) with a mean of 12.032.

The marginal ridge component showed a range of points deducted for the four examiners of (1 to 25.5), where Examiner 4 showed the lowest points deduction with a range of (1 to 25) with a mean of 5.839. On the other hand, Examiner 1 showed the highest points deduction for marginal ridge component with a range of (5 to 25.5) with a mean of 10.725.

The occlusal relationship component showed a range of points deducted for the four examiners of (1 to 32), where Examiner 3 showed the lowest points deduction with a range of (1 to 22) with a mean of 10.161. On the other hand, Examiner 1 showed the highest points deduction for the occlusal relationship component with a range of (7.5 to 25) with a mean of 17.241.

The occlusal contacts component showed a range of points deducted for the four examiners of (7 to 33), where Examiner 1 showed the lowest points deduction with a range of (7 to 29) with a mean of 19.241. Examiner 3 showed the highest points deduction for occlusal contacts component with a range of (14 to 33) with a mean of 26.161.

5.1.3 The difference in points between digital and plaster models

The ABO OGS components and the total score difference in points deduction between digital and plaster models are shown in Table 5.3. The difference in mean total score between digital and plaster models was high; Examiner 2 showed the highest with a mean difference of 102.35 and the lowest was Examiner 4 with a mean difference of 67.935.

The difference between digital and plaster for the buccolingual inclination component showed the lowest correlation between digital and plaster models. Examiner 1 showed the highest mean difference of 2.967. While Examiners 3 and 4 showed that

plaster was higher than digital: Examiner 3 = -0.580 and Examiner 4 = -0.290.

The difference in mean overjet between digital and plaster models was high for all examiners. Examiner 4 showed the highest with a mean difference of 22.19; on the other hand, Examiner 1 showed the lowest points deducted with a mean difference of 20.612.

Interproximal contacts component showed the highest mean difference between digital and plaster models, as Examiner 3 showed the highest and Examiners 1 and 2 also showed a high mean difference while Examiner 4 showed the lowest between examiners with very low mean difference compared to the other 3 examiners.

The difference in mean between digital and plaster models for the occlusal contacts component showed a high difference for all examiners. Examiner 4 showed the highest points deducted with a mean difference of 23.322; on the other hand, Examiner 1 showed the lowest points deducted with a mean difference of 15.596.

Examiner 2 showed the highest points deducted in the occlusal relationship component with a mean difference of 20.903; on the other hand, Examiner 3 showed the lowest points deducted with a mean difference of 6.419.

Examiner 4 showed the highest points deducted in the alignment component with a mean difference of 8.741; on the other hand, Examiner 2 showed the lowest points deducted with a mean difference of 4.322. Examiner 1 showed the highest points deducted in the marginal ridge component with a mean difference of 7.580; on the other hand, Examiner 4 showed the lowest points deducted with a mean difference of 2.774.

Table 5.1 Mean, standard deviation and range for the four examiners when measuring the seven components of the ABO OGS and total scores using 31 plaster models

	Examiner 1				Examiner 2				Examiner 3				Examiner 4			
	Mean	SD	Range		Mean	SD	Range		Mean	SD	Range		Mean	SD	Range	
			High	Low			High	Low			High	Low			High	Low
Alignment	3.532	1.910	9.00	1.00	3.35	1.761	8.0	1.0	3.387	1.605	7.0	1.0	3.290	1.918	9.0	0.00
Marginal Ridge	3.145	1.649	7.50	0.00	3.161	1.968	9.0	0.00	3.129	1.944	9.0	0.00	3.065	1.878	8.0	0.00
Buccolingual Inclination	4.612	2.319	10.5	0.00	4.645	2.677	12.0	0.00	4.710	2.830	12.0	0.00	4.710	3.002	12.0	0.00
Occlusal Relationship	3.645	2.840	14.00	0.00	3.774	3.412	17.0	0.00	3.742	3.415	17.0	0.00	3.548	3.548	17.0	0.00
Occlusal Contacts	2.919	3.423	14.50	0.00	2.839	3.821	15.0	0.00	2.839	3.697	13.0	0.00	2.516	3.749	15.0	0.00
Overjet	3.467	2.546	8.00	0.00	3.387	2.564	8.0	0.00	3.323	2.677	8.0	0.00	3.774	3.232	11.0	0.00
Interproximal Contacts	0.838	0.799	3.00	0.00	0.581	0.719	2.0	0.00	0.613	0.803	2.0	0.00	0.226	0.497	2.0	0.00
Total Score	22.143	7.540	48.00	17.5	21.742	7.966	45.0	10.0	21.742	8.660	46.0	9.0	21.387	8.428	46.0	9.00

Table 5.2 Mean, standard deviation and range for the four examiners when measuring the seven components of the ABO OGS and total scores using 31 digital models

	Examiner 1				Examiner 2				Examiner 3				Examiner 4			
	Mean	SD	Range		Mean	SD	Range		Mean	SD	Range		Mean	SD	Range	
			High	Low			High	Low			High	Low			High	Low
Alignment	11.225	2.351	15.50	7.00	7.677	4.237	19.0	2.0	9.226	6.940	31.0	1.0	12.032	6.959	29.0	1.0
Marginal Ridge	10.725	5.024	25.50	5.00	9.097	6.467	24.0	1.0	6.161	2.647	14.0	2.0	5.839	4.132	25.0	1.0
Buccolingual Inclination	7.580	3.071	15.00	3.00	6.290	3.7967	19.0	2.0	4.129	2.604	12.0	.0	4.419	1.945	9.0	.0
Occlusal Relationship	17.241	5.113	25.00	7.50	24.677	4.949	32.0	12.0	10.161	5.020	22.0	1.0	11.871	3.792	20.0	5.0
Occlusal Contacts	19.241	6.434	29.00	7.00	24.065	4.992	33.0	14.0	26.161	4.810	33.0	14.0	22.258	4.381	33.0	15.0
Overjet	24.080	4.229	33.00	15.00	24.871	4.910	34.0	11.0	25.097	2.820	29.0	19.0	25.903	3.037	28.0	17.0
Interproximal Contacts	24.016	4.225	31.00	14.50	27.419	4.537	35.0	18.0	27.871	4.514	41.0	18.0	6.548	4.780	25.0	2.0
Total Score	114.27	10.21	138.0	98.00	124.097	12.918	148.0	90.0	107.968	13.382	133.0	82.0	89.323	14.549	133.0	66.0

Table 5.3 Mean, standard deviation and range of the differences between plaster and digital for the four examiners

	Examiner 1				Examiner 2				Examiner 3				Examiner 4			
	Mean	SD	Range		Mean	SD	Range		Mean	SD	Range		Mean	SD	Range	
			High	Low			High	Low			High	Low			High	Low
Alignment	7.693	3.182	13.50	2.00	4.322	4.657	16.00	-3.00	5.837	6.976	26.00	-3.00	8.741	6.592	23.00	-2.00
Marginal Ridge	7.580	5.094	23.00	-1.00	5.935	7.042	23.00	-3.00	3.033	3.027	12.00	-1.00	2.774	4.302	23.00	-1.00
Buccolingual Inclination	2.967	3.612	11.00	-4.50	1.645	5.288	19.00	-6.00	-0.580	4.088	12.00	-10.00	-0.290	2.648	5.00	-9.00
Occlusal Relationship	13.596	5.724	24.00	2.00	20.903	5.230	29.00	10.00	6.419	5.439	15.00	-7.00	8.332	3.134	15.00	2.00
Occlusal Contacts	15.596	6.667	26.50	.00	21.221	6.286	32.00	5.00	23.322	6.062	33.00	6.00	19.741	5.247	33.00	6.00
Overjet	20.612	4.600	31.00	8.00	21.483	6.082	32.00	3.00	21.774	3.343	28.00	16.00	22.19	4.326	28.00	8.00
Interproximal Contacts	23.177	4.358	31.00	12.50	26.838	4.509	34.00	17.00	27.258	4.735	41.00	17.00	6.323	4.700	24.00	1.00
Total Score	86.371	13.014	110.50	61.00	102.35	16.491	138.0	61.00	86.225	16.115	118.00	51.00	67.935	15.524	104.0	36.0

5.2 Examiner reliability

Results for the Pearson correlation coefficient for the intra-examiner and inter-examiner reliability are shown in Tables 5.4, 5.5 and 5.6.

Table 5.4 Pearson correlation coefficients for comparisons of repeated measurements for Examiner 1 of plaster and digital models

Comparison	Alignment	Marginal Ridge	Buccolingual Inclination	Occlusal Contacts	Occlusal Relationship	Overjet	Inter proximal Contacts	Total Score
Examiner1 (Plaster 1) versus Examiner1 (Plaster 2)	R=0.885** P=0.000	R=0.932** P=0.000	R=0.938** P=0.000	R=0.915** P=0.000	R=0.976** P=0.000	R=0.982** P=0.000	R=0.771** P=0.000	R=0.915** P=0.000
Examiner1 (Digita1 1) versus Examiner1 (Digital 2)	R=0.352 P=0.052	R=0.561** P=0.001	R=0.496** P=0.005	R=0.846** P=0.000	R=0.883** P=0.000	R=0.216 P=0.244	R=0.555** P=0.001	R=0.583** P=0.001

Plaster 1 = Examiner 1 first measurement of the 31 plaster models

Plaster 2 =Examiner 1 second measurement of the 31 plaster models

Digital 1= Examiner 1 first measurement of the 31 digital models

Digital 2= Examiner 2 second measurement the 31 digital models

Examiner 1 measured the ABO OGS components twice for intra-examiner reliability and the data was analysed using a Pearson correlation test, for both

plaster and digital models. High correlation was found for Examiner 1's repeated measures (R=0.915), suggesting high intra-examiner reliability for the total ABO OGS score using plaster. This correlation was found to be statistically significant (P=0.00). The lowest correlation found was the interproximal contacts component (R=0.771). Moderate correlation was found for the total score for digital models (R=0.583).This correlation was found to be statistically significant (p=0.001). Overjet showed the lowest correlation for digital models.

5.2.1 Examiners' reliability for plaster models

Table 5.5 Pearson Correlation coefficients and P values for comparisons between each examiner and record type for each ABO OGS component and total score to show inter-examiner reliability for plaster models

Comparison	Alignment	Marginal Ridge	Buccolingual Inclination	Occlusal Relationship	Occlusal Contacts	Overjet	Inter-proximal Contacts	Total Score
Examiner 1 versus Examiner 2	R=0.938** P=0.000	R=0.896** P=0.000	R=0.970** P=0.000	R=0.975** P=0.000	R=0.971** P=0.000	R=0.992** P=0.000	R=0.747** P=0.000	R=0.960** P=0.000
Examiner 1 versus Examiner 3	R=0.860** P=0.000	R=0.919** P=0.000	R=0.955** P=0.000	R=0.964** P=0.000	R=0.961** P=0.000	R=0.978** P=0.000	R=0.626** P=0.000	R=0.927** P=0.000
Examiner 1 versus Examiner 4	R =0.356* P = 0.049	R=0.243 P =0.187	R =0.440* P = 0.013	R = 0.627** P = 0.000	R =0.167 P =0.369	R =0.099 P = 0.597	R =0.020 P =0.917	R=0.814** P =0.000
Examiner 2 versus Examiner 3	R=0.951** P=0.000	R=0.908** P=0.000	R=0.949** P=0.000	R=0.984** P=0.000	R=0.982** P=0.000	R=0.977** P=0.000	R=0.805** P=0.000	R=0.970** P=0.000
Examiner 2 versus Examiner 4	R=0.965** P=0.000	R=0.908** P=0.000	R=0.940** P=0.000	R=0.977** P=0.000	R=0.832** P=0.000	R=0.863** P=0.000	R=0.460** P=0.009	R=0.877** P=0.000
Examiner 3 versus Examiner 4	R=0.904** P=0.000	R=0.937** P=0.000	R=0.982** P=0.000	R=0.983** P=0.000	R=0.893** P=0.000	R=0.891** P=0.000	R=0.393* P=0.029	R=0.885** P=0.000

Table 5.5 shows the inter-examiner reliability for plaster models for the four examiners. The highest correlations were found between Examiners 1, 2 and 3; however, Examiner 4 had the lowest correlation with the other examiners.

5.2.2 Examiners' reliability for digital models

Table 5.6 Pearson correlation coefficients and P values for comparisons between each examiner and record type for each ABO OGS component and total score to show inter-examiner reliability for digital models.

Comparison	Alignment	Marginal Ridge	Buccolingual Inclination	Occlusal Relationship	Occlusal Contacts	Overjet	Inter-proximal Contacts	Total Score
Examiner 1 versus Examiner 2	R = 0.023 P = 0.904	R = -0.098 P = 0.599	R = 0.057 P = 0.763	R = 0.071 P = 0.704	R =-0.018 P = 0.925	R = -0.314 P = 0.085	R = 0.052 P = 0.782	R = 0.181 P = 0.329
Examiner 1 versus Examiner 3	R = 0.289 P = 0.115	R = 0.100 P = 0.593	R = -0.081 P = 0.667	R = -0.179 P = 0.334	R = 0.148 P = 0.426	R = -0.180 P = 0.334	R = -0.114 P = 0.540	R = 0.057 P = 0.762
Examiner 1 versus Examiner 4	R = -0.113 P = 0.543	R = 0.035 P = 0.853	R = -0.106 P = 0.569	R= -0.296 P = 0.106	R =-0.555** P = 0.001	R = -0.010 P = 0.958	R = 0.181 P = 0.330	R = -0.039 P = 0.834
Examiner 2 versus Examiner 3	R = -0.188 P = 0.312	R = -0.028 P = 0.880	R = 0.141 P = 0.449	R = -0.001 P = 0.998	R = 0.209 P = 0.259	R = -0.100 P = 0.592	R = -0.119 P = 0.523	R = -0.238 P = 0.197

Examiner 3 versus Examiner 4	R = 0.474** P = 0.007	R =0.569** P = 0.001	R=0.410* P = 0.022	R = 0.346 P = 0.057	R = 0.121 P = 0.516	R = 0.332 P = 0.068	R = -0.282 P = 0.124	R = 0.415* P = 0.020
Examiner 2 versus Examiner 4	R = -0.055 P = 0.769	R = -0.109 P = 0.559	R = -0.243 P = 0.188	R = -0.059 P = 0.752	R = 0.031 P = 0.868	R = -0.215 P = 0.244	R = 0.169 P = 0.364	R = -0.469** P = 0.008

A very low correlation was found between examiners in the digital models. The only moderate correlations were found between Examiners 3 and 4 for the total score with a correlation of 0.415 and significance of *P* = 0.020 (Table 5.6).

5.3 Validation of the new software

As Examiner 1 was found to have high intra-examiner reliability for plaster models (total score R = 0.915) and moderate reliability for digital models (R = 0.583) (Table 5.4) it was decided to use Examiner 1's data as a representative for the statistical comparison used in the current study.

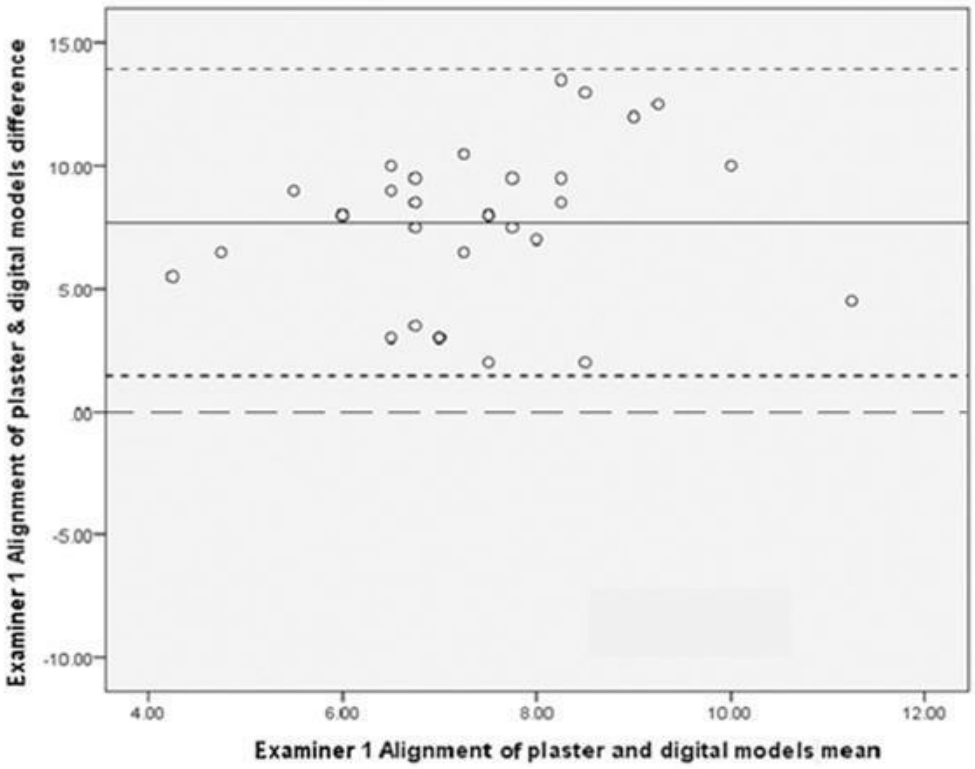


Figure 5.1 Bland-Altman scatter plot for alignment for Examiner1

Figure 5.1 shows the Bland and Altman scatter plot for the alignment component of ABO OGS using the mean of the repeated measurements. The mean difference was 7.69 and 95% limits of agreement was 1.45 to 13.93. This indicates no agreement between different methods. All digital measurements are higher than plaster models. All of the plots are above the zero horizontal line; this indicates that there is no agreement.

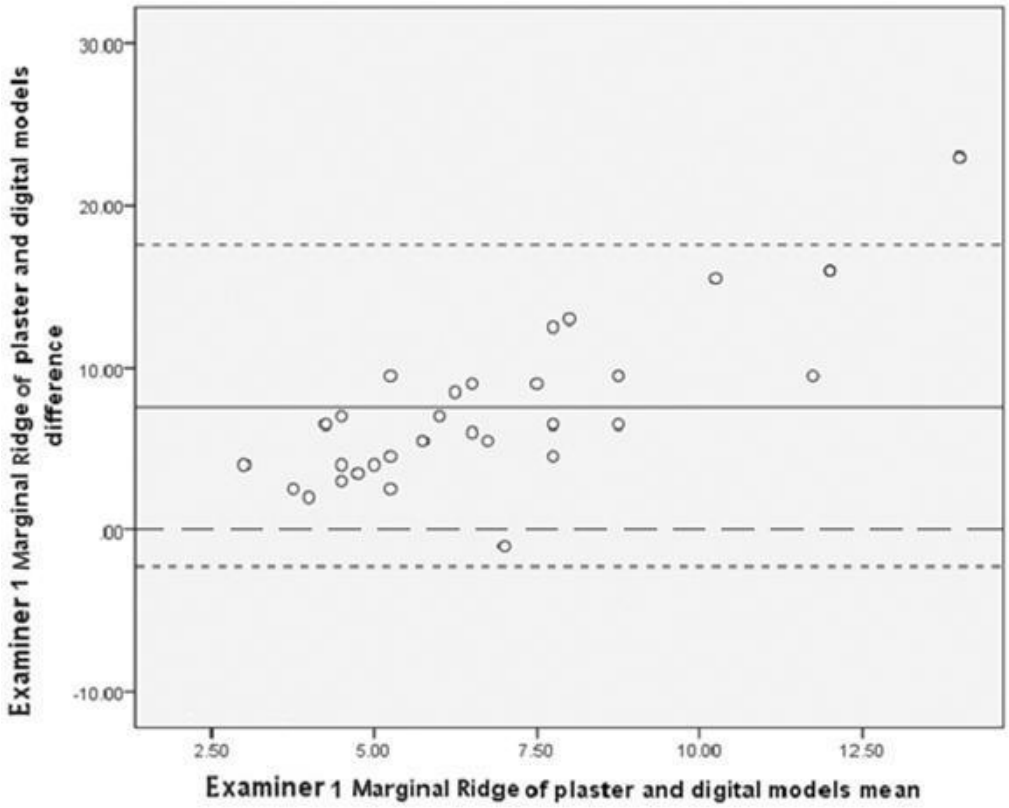


Figure 5.2 Bland-Altman scatter plot for the marginal ridge for Examiner 1

Figure 5.2 shows the Bland and Altman scatter plot for the marginal ridge component of ABO OGS using the mean of the repeated measurements. The mean difference was 7.56 and 95% limits of agreement was 17.56 to -2.28. This indicates very low agreement between different methods.

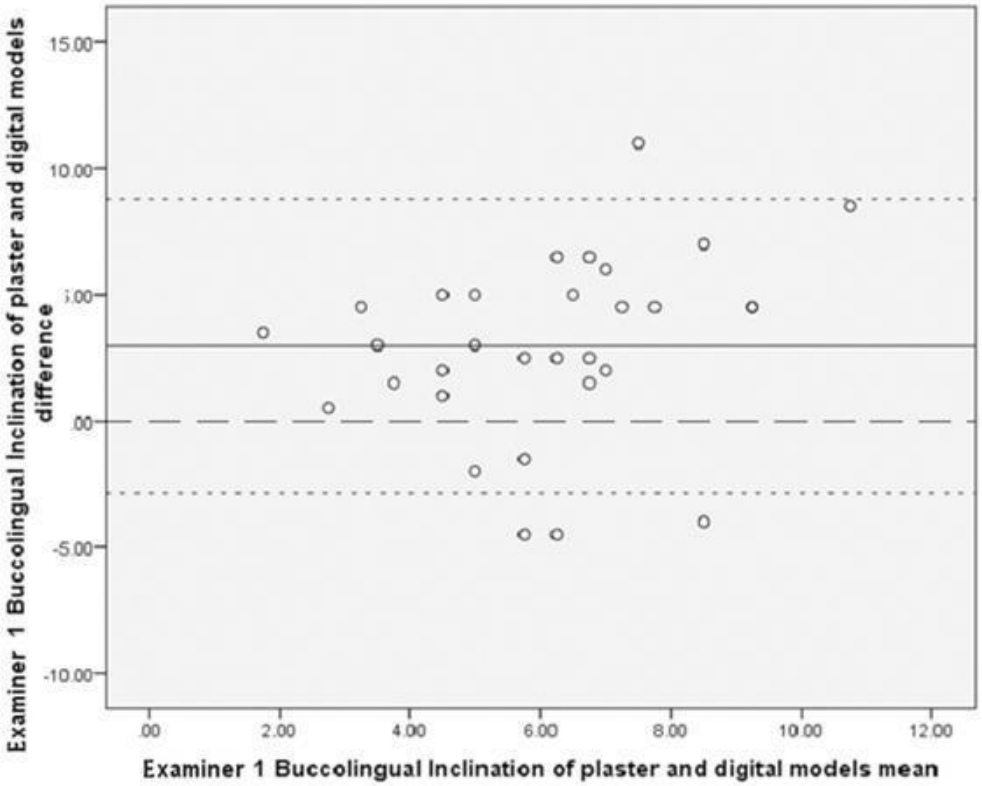


Figure 5.3 Bland-Altman scatter plot for the buccolingual inclination for Examiner 1

Figure 5.3 shows the Bland and Altman scatter plot for the buccolingual inclination component of ABO OGS using the mean of the repeated measurements. The mean difference was 2.96 and 95% limits of agreement was 8.78 to -2.84. This indicates no agreement between the different methods. All digital measurements are higher than the plaster models. The plots sometimes are less than zero, which indicates that digital measurements are sometimes less than plaster.

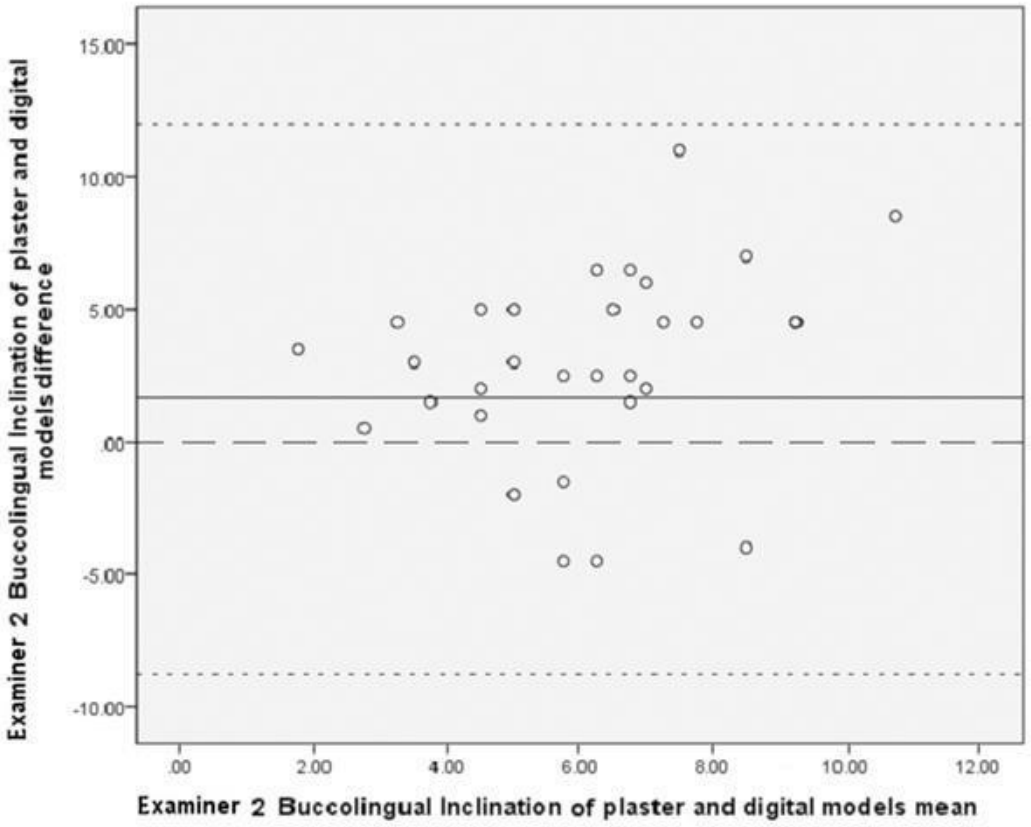


Figure 5.4 Bland-Altman scatter plot for the buccolingual inclination for Examiner 2

Figure 5.4 shows the Bland and Altman scatter plot for the buccolingual inclination component of ABO OGS using the mean of the repeated measurements. The mean difference was 1.67, indicating acceptable agreement between the two methods. The 95% limits of agreement were 12.09 to 8.78, which shows a wide range.

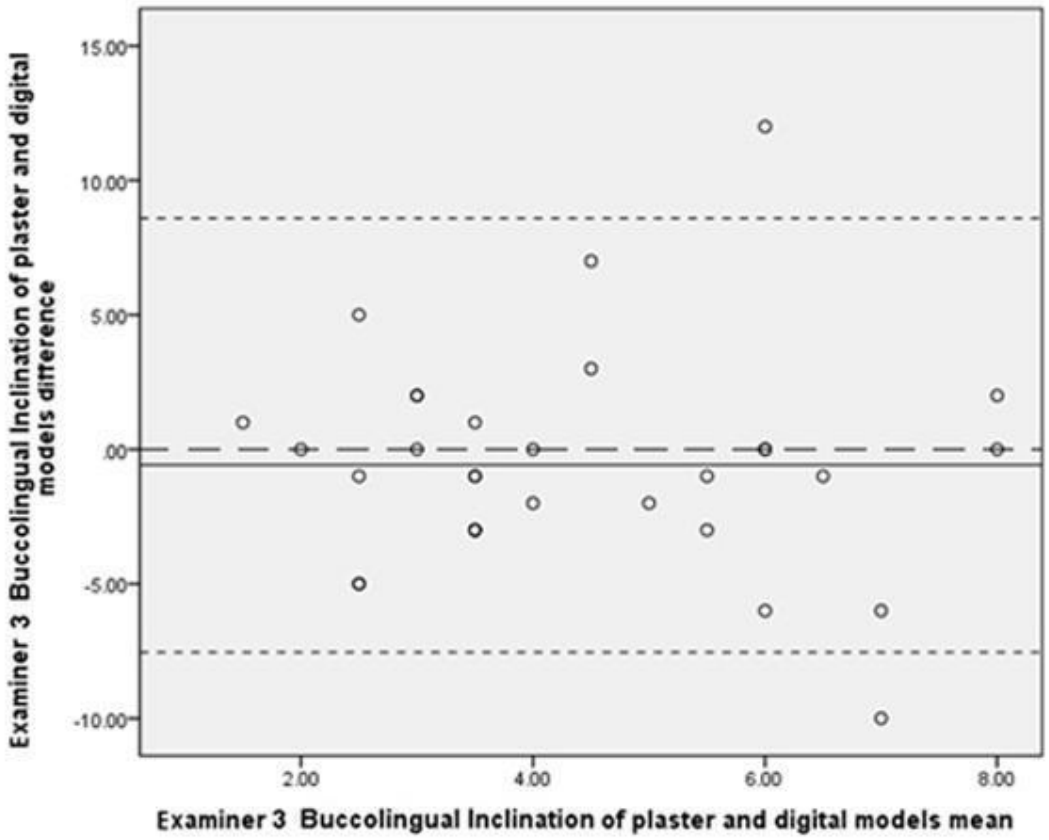


Figure 5.5 Bland-Altman scatter plot for the buccolingual inclination for Examiner 3

Figure 5.5 shows the Bland and Altman scatter plot for the buccolingual inclination component of ABO for Examiner 3. The mean and limits of agreement for Examiner 3 show mean difference of alignment = 0.5806, which indicates a clinically acceptable agreement. The limits of agreement (95% limits of agreement) were 8.59 to 7.42, which showed a wide range.

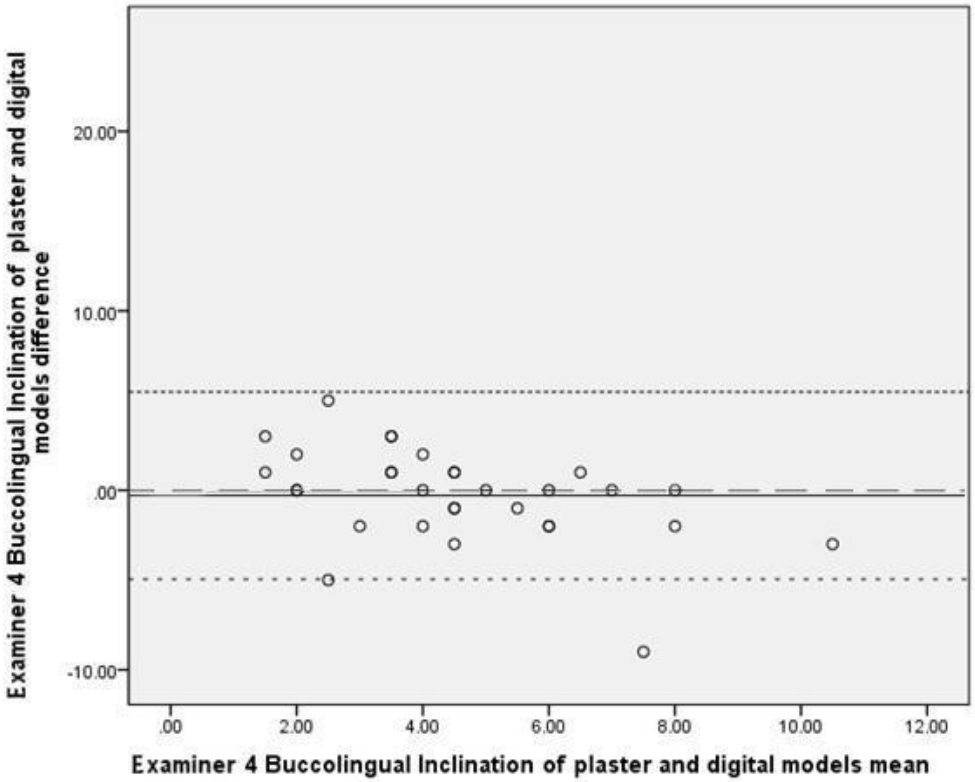


Figure 5.6 Bland-Altman scatter plot for the buccolingual inclination for Examiner 4

Figure 5.6 shows the Bland and Altman scatter plot for the buccolingual inclination component of ABO for Examiner 4 using the mean of the repeated measurements for Examiner 4. The mean difference was -0.290 and 95% limits of agreement was 5.48 to -4.89. This indicates reasonable agreement between the different methods.

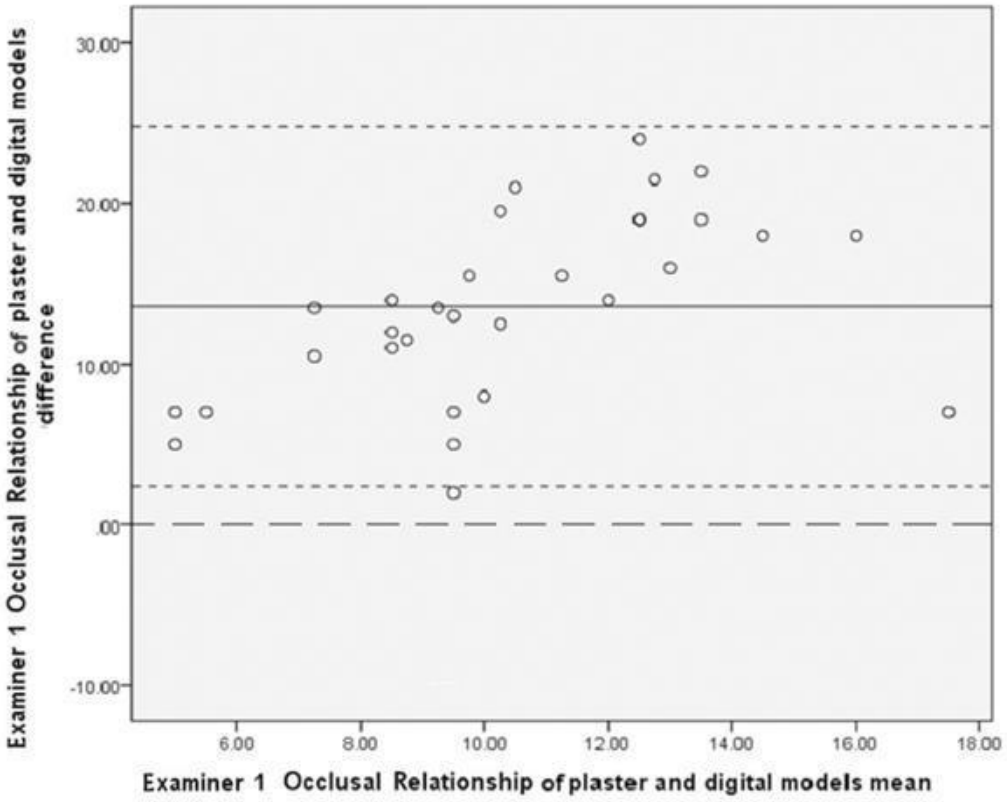


Figure 5.7 Bland-Altman scatter plot for the occlusal relationship for Examiner 1

Figure 5.7 shows the Bland and Altman scatter plot for the occlusal relationship component of ABO using the mean of the repeated measurements. The mean difference was 5.72 and 95% limits of agreement was 24.81 to 2.37. This indicates no agreement between the different methods. All digital measurements are higher than plaster models.

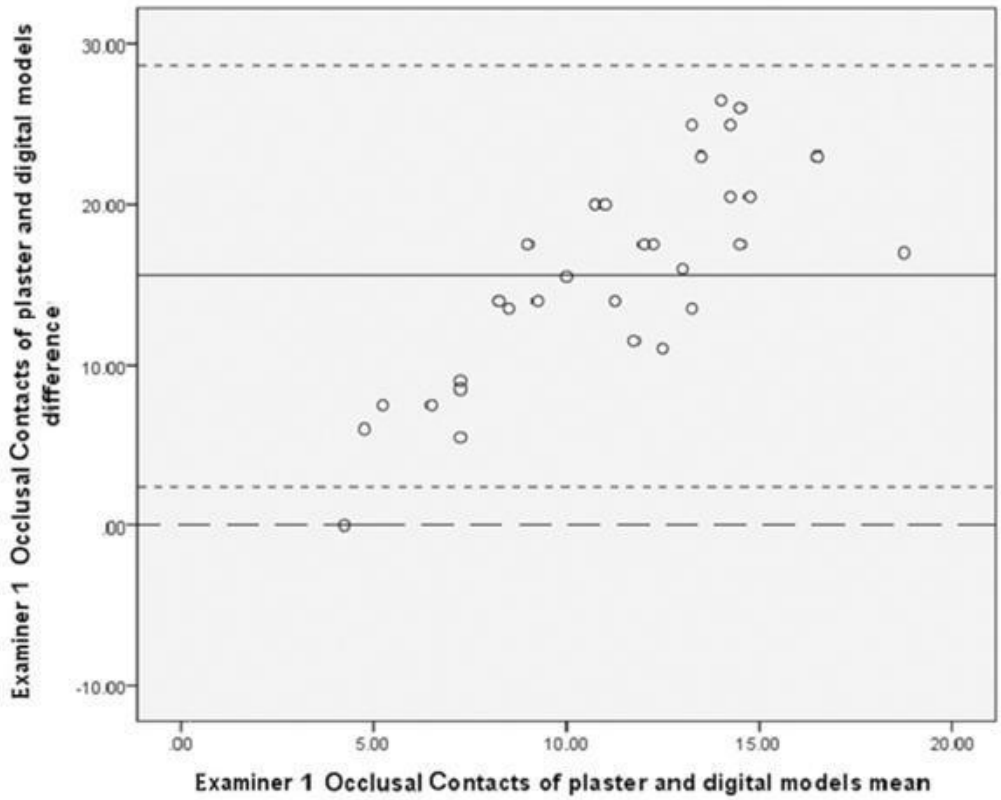


Figure 5.8 Bland-Altman scatter plot for the occlusal contacts for Examiner 1

Figure 5.8 shows the Bland and Altman scatter plot for the occlusal contacts component of ABO using the mean of the repeated measurements. The mean difference was 15.59 and 95% limits of agreement was 28.66 to 2.37. This difference is considered to be a clinically significant difference suggesting no agreement between the two different methods. Generally, the digital measurements are higher than the plaster models.

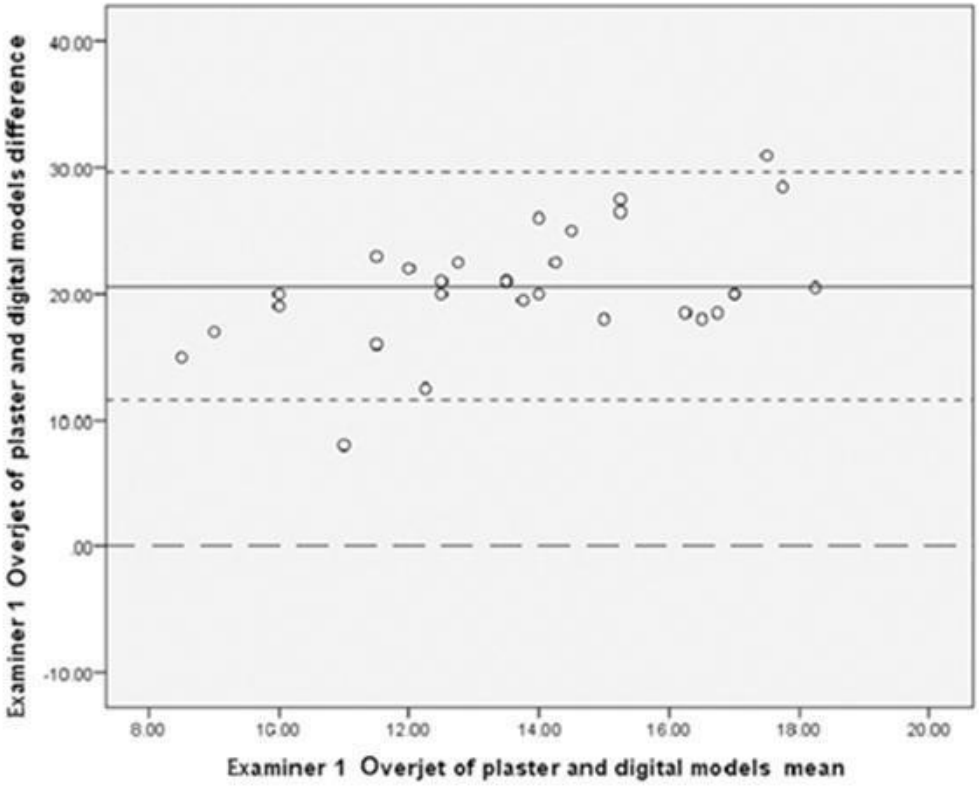


Figure 5.9 Bland-Altman scatter plot for the overjet for Examiner 1

Figure 5.9 shows the Bland and Altman scatter plot for the overjet component of ABO using the mean of the repeated measurements. The mean difference was 20.61 and 95% limits of agreement= 29.62 to 11.59. This difference was found to be clinically significant, suggesting no agreement between different methods and that the digital measurements are higher than the plaster models.

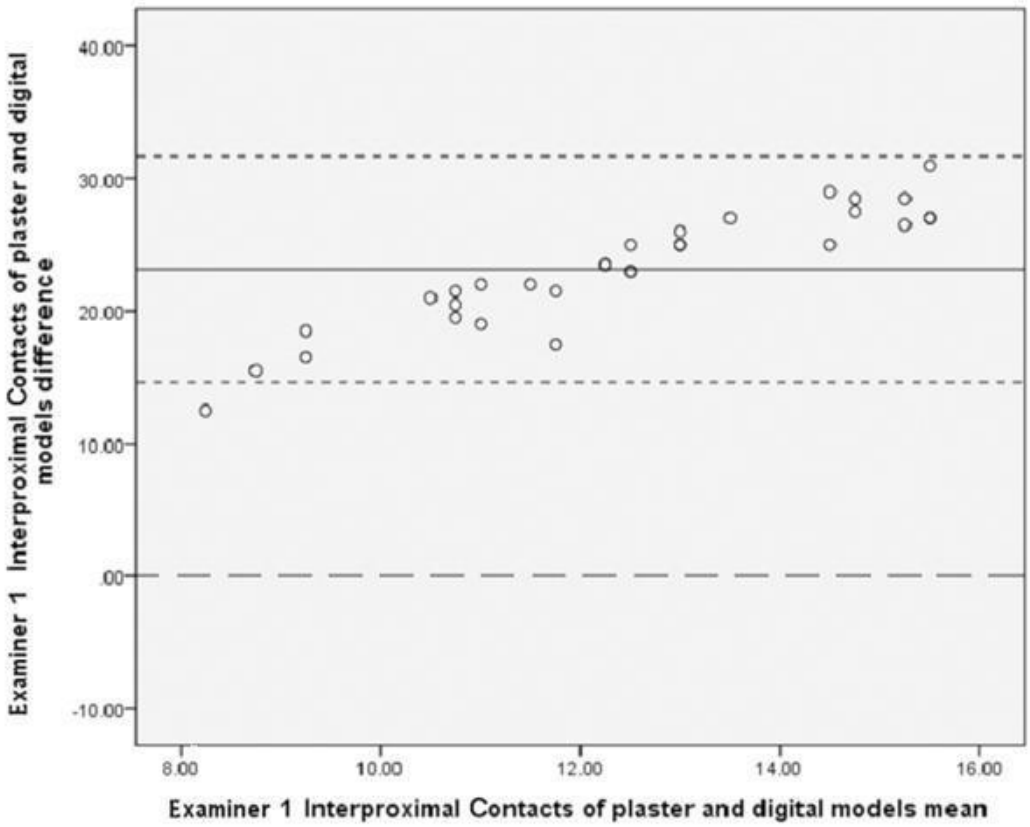


Figure 5.10 Bland-Altman scatter plot for the inter-proximal contacts for Examiner 1

Figure 5.10 shows the Bland and Altman scatter plot for the inter-proximal contacts component of ABO for Examiner 1 using the mean of the repeated measurements. The mean difference was 23.17 and 95% limits of agreement = 31.69 to 14.65. This difference was found to be clinically significant, suggesting no agreement between the different methods. Generally, the digital measurements are higher than the plaster models.

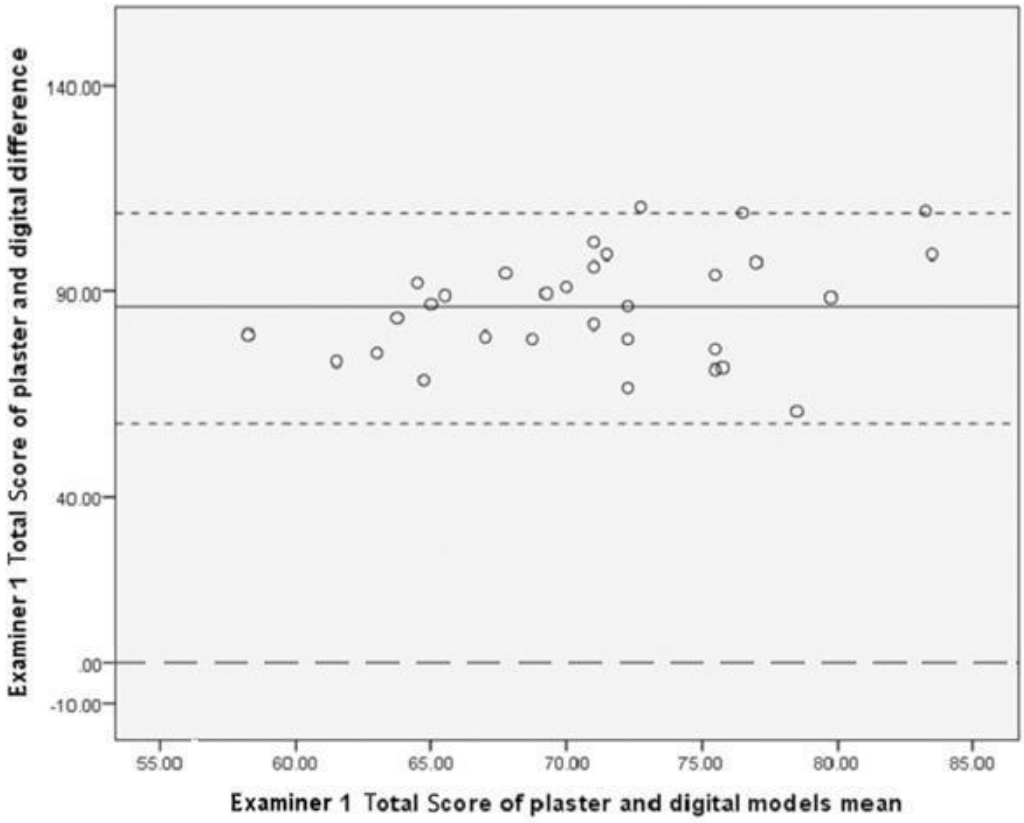


Figure 5.11 Bland-Altman scatter plot for the total score for Examiner 1

Figure 5.11 shows the Bland and Altman scatter plot for the total score of ABO OGS for Examiner 1 using the mean of the repeated measurements. The mean difference was 83.36 and 95% limits of agreement = 108 to 57.86. This difference was found to be clinically significant, suggesting no agreement between the different methods. Generally, the digital measurements were higher than the plaster models.

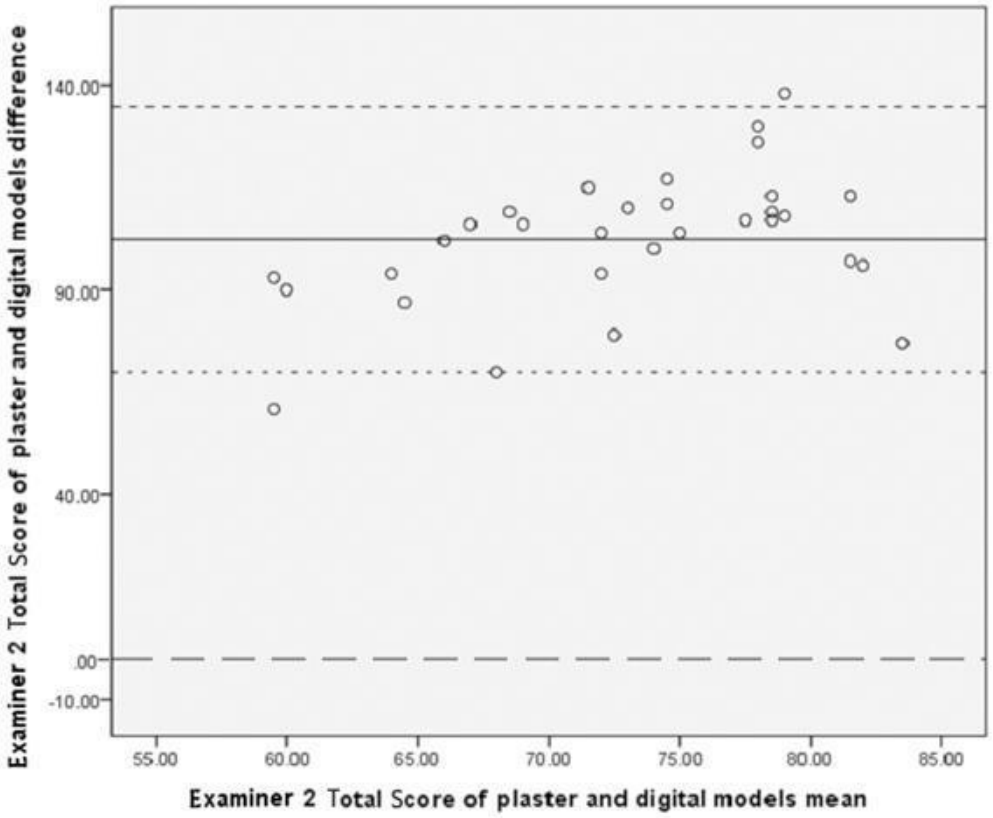


Figure 5.12 Bland-Altman scatter plot for the total score for Examiner 2

Figure 5.12 shows the Bland and Altman scatter plot for the total score of ABO OGS for Examiner 2 using the mean of the repeated measurements. The mean difference was 102.35 and 95% limits of agreement = 134.67 to 70. This difference was found to be clinically significant, suggesting no agreement between the different methods. Generally the digital measurements were higher than the plaster models.

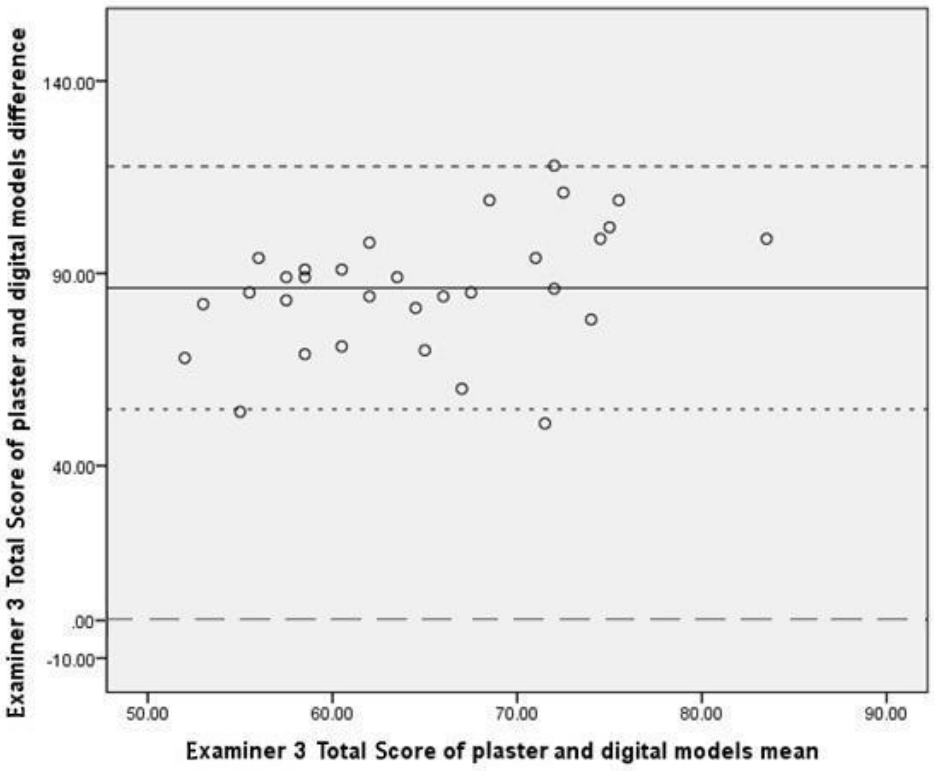


Figure 5.13 Bland-Altman scatter plot for the total score for Examiner 3

Figure 5.13 shows the Bland and Altman scatter plot for the total score of ABO OGS for Examiner 3 using the mean of the repeated measurements. The mean difference was 86.22 and 95% limits of agreement= 117.8 to 54.64. This difference was found to be clinically significant, suggesting no agreement between the different methods. Generally the digital measurements were higher than the plaster models.

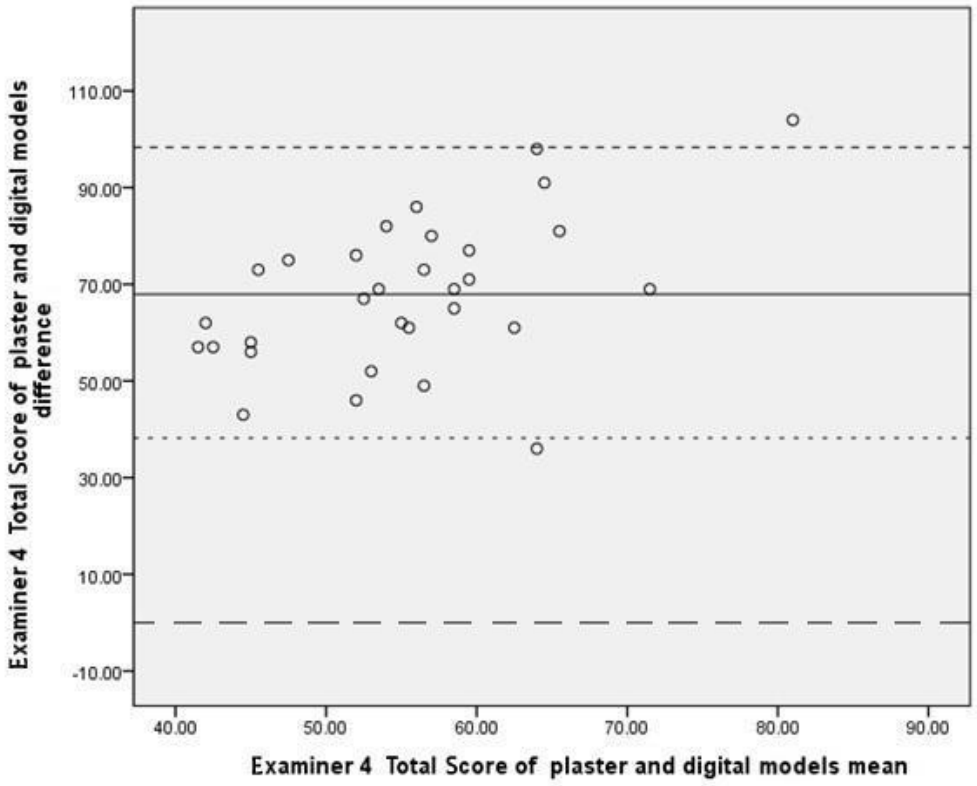


Figure 5.14 Bland-Altman scatter plot for the total score for Examiner 4

Figure 5.14 shows the Bland and Altman scatter plot for the total score of ABO OGS for Examiner 4 using the mean of the repeated measurements. The mean difference was 67.93 and 95% limits of agreement= 98.36 to 37.50. This difference was found to be clinically significant, suggesting no agreement between the different methods. Generally the digital measurements were higher than the plaster models.

Table 5.7 Mean difference and limit of agreements of plaster and digital models measurements

	Examiner 1			Examiner 2			Examiner 3			Examiner 4		
	Mean Of Difference	95% Limits Of Agreement		Mean Of Difference	95% Limits Of Agreement		Mean Of Difference	95% Limits Of Agreement		Mean Of Difference	95% Limits Of Agreement	
		High	Low		High	Low		High	Low		High	Low
Alignment	7.69	13.93	1.45	4.32	13.45	-4.80	8.83	19.51	-7.83	8.74	21.66	-4.18
Marginal Ridge	7.56	17.56	-2.28	5.93	19.73	-7.86	3.03	8.86	-2.90	2.77	11.20	-5.66
Buccolingual Inclination	2.96	8.78	-2.84	1.64	12.09	-8.78	-.5806	8.59	-7.42	-0.290	5.48	-4.89
Occlusal Relationship	5.72	24.81	2.376	20.90	31.15	10.65	6.41	17.07	-4.2	8.32	14.46	2.17
Occlusal Contacts	15.59	28.66	2.37	21.22	33.54	8.9	23.32	35.20	11.43	19.74	30.026	9.45
Overjet	20.61	29.62	11.59	21.48	33.4	9.56	21.77	28.32	15.22	22.12	30.60	13.64
Interproximal Contacts	23.17	31.69	14.65	26.83	33.66	18.00	27.25	35.51	17.99	6.32	15.53	-2.89
Total Score	86.37	108.87	57.86	102.35	134.67	70.02	86.22	117.8	54.64	67.93	98.36	37.50

5.4 Distribution of the scoring result of the ABO OGS for the sample using the conventional plaster model and digital models

Table 5.8 distribution of the scoring result of the ABO OGS (Examiners 1-4) for the sample using the conventional plaster model and the ABO software

Examiner 1	**Passed number**	**Passed %**	**Borderline number**	**Borderline %**	**Fail number**	**Fail %**	**Chi-Square Test**
Conventional (Plaster models)	4	12.9%	16	51.6%	11	35.5%	**Chi-Square:** 7.032[a] **Df :** 2 **Asymp. Sig.:** 0.030
ABO software (Digital models)	0	0	0	0	31	100%	
Examiner 2	**Passed number**	**Passed %**	**Borderline number**	**Borderline %**	**Fail number**	**Fail %**	
Conventional (Plaster models)	13	41.9%	13	41.9%	5	16.1%	**Chi-Square:** 4.129[a] **Df:** 2 **Asymp. Sig.:** 0.127
ABO software (Digital models)	0	0	0	0	31	100%	
Examiner 3	**Passed number**	**Passed %**	**Borderline number**	**Borderline %**	**Fail number**	**Fail %**	
Conventional (Plaster models)	12	38.7%	11	35.5%	8	25.8%	**Chi-Square:** 0.839[a] **Df:** 2 **Asymp. Sig.:** 0.657
ABO software (Digital models)	0	0	0	0	31	100%	
Examiner 4	**Passed number**	**Passed %**	**Borderline number**	**Borderline %**	**Fail number**	**Fail %**	
Conventional (Plaster models)	13	41.9%	14	45.2%	4	12.9%	**Chi-Square:** 5.871[a] **Df:** 2 **Asymp. Sig.:** 0.053
ABO software (Digital models)	0	0	0	0	31	100%	

Table 5.8 shows the distribution of the scoring result of the ABO OGS (Examiners 1-4) for the sample using the conventional plaster model and the ABO software. According to the ABO examination board guidelines a score of less than 30 points is a pass, between 20 and

30 is a boarder pass (grey zone) and score more than 30 is a clear fail. It was found that the four examiners scored the entire sample (100%) as failed to pass using ABO software, while using the conventional plaster methods the fail rate ranged between 12.9% and 35.5 % of the whole sample. This difference in the fail rate for the same sample between the ABO software and the conventional method was found to be statistically significant ($P<0.05$) for the four the examiners.

5.5 Time for ABO scoring

5.5.1 Time for ABO OGS scoring digital models using the new software

Times to measure the digital models using the new software are shown in Table 5.9.

Table 5.9 ANOVA test to compare the time taken to score thirty one digital models.

Examiner	Number of models	Mean time (minutes)	SD	Mean square	F	Significance (*P* value)
Examiner 1	31	30.94	3.11	178.954	10.528	0.063
Examiner 2	31	35.94	4.28			
Examiner 3	31	33.65	4.13			
Examiner 4	31	36.03	4.90			

Mean difference is significant at $P<0.05$

A Shapiro-Wilk test was used to test the normality of the data and was found to be statistically insignificant ($P>0.05$). An ANOVA test was used to investigate the difference in mean time (minutes) between the four different examiners in scoring the digital models (Table 5.9). There was no statistically significant difference found between the four examiners ($P = 0.063$).

5.5.2 Time for ABO scoring using plaster models

Mean time to measure the plaster models using are shown in Table 5.10. An ANOVA test was used to investigate the difference in mean time (minutes) between the four different examiners in ABO scoring the plaster models (Table 5.10). There was no statistically significant difference found between the four examiners (P = 0.635).

Table 5.10 ANOVA test to compare the time taken by different examiners to score the plaster models.

Examiner	Number of models	Mean time (minutes)	SD	Mean square	F	Significance (P value)
Examiner 1	31	11.1935	2.36	621.642	66.004	0.635
Examiner 2	31	11.9032	3.48			
Examiner 3	31	11.1935	2.36			
Examiner 4	31	20.4194	3.83			

Mean difference is significant at P<0.05

5.5.3 Comparison between times for ABO OGS scoring using plaster and digital models

The 4 examiners' scoring times (each scored 31 models) were added together to represent a total of (31x4) 124 scoring attempts by the four examiners (Table 5.11). The mean time taken by the examiners to ABO score digital and plaster models was 34.14 SD +/- 4.57 minutes and 13.7 SD +/-4.50 minutes, respectively. Paired t-test was used to investigate the difference in mean time for ABO scoring between plaster and digital models for the total sample (Table 5.11). There was a statistically significant (P = 0.000) increase in time for ABO scoring in the digital model compared with the plaster models (mean difference 20.43 minutes, CI -21.64 to 19.18).

Table 5.11 Paired t-test to compare the time taken for ABO OGS scoring between the plaster and digital models

	Total number of the 4 examiners scoring (each scored 31 models)	Mean time (minutes)	SD	Mean difference	95% Confidence Intervals (CI)		Sig
					Lower	Upper	
Plaster models	124	13.71	4.93	20.43	-21.64	-19.18	0.000
Digital models	124	34.14	4.5				

Mean difference is significant at $P<0.05$

Chapter 6: Discussion

6.1 Descriptive analysis

6.1.1 Study sample size

Thirty one post-treatment study models were used in this investigation. All cases had completed fixed appliance orthodontic treatment in both mandibular and maxillary arches. The sample size used in the current investigation is similar to that reported (from 24-36 models) by the three studies conducted previously to investigate the validity of different ABO OGS digital models software systems (Costalos et al., 2005, Okunami et al., 2007, Hildebrand et al., 2008).

Although the sample size in the current study was comparable to the previous similar studies, prior sample size calculation was not done. To ensure that the sample size was sufficient to undertake the statistical analysis, sample size calculation was done retrospectively using the results of the most recent study by Hildebrand et al. (2008). A sample size of 16 study models in each group would have 80% power to detect a difference of 5 points of the total score, assuming that the common standard deviation is 4.9 using a 0.05 significant difference level (http://www.statisticalsolutions.net/pss_calc.php). This may suggest that the study had enough power to detect variation between the two methods.

6.1.2 Description of the ABO OGS scores

According to the ABO specifications, a total deduction of 20 points is a passing score, and a deduction of 30 points, is a failing score for assessing the study models for passing the ABO examination. The range between 20 and 30 points deducted is a grey zone. For the current study, the mean total deductions score among the examiners for plaster models ranged from was as low as 21.387 SD +/- 8.428 to as high as 22.143 SD +/- 7.54 (Table 5.1), which is in the grey zone. However, for the digital models the mean total deductions score among the examiners ranged from as low as 89.323 SD +/- 14.549 to as high as 124.097 SD +/- 12.918 points deduction, which cannot be accepted to pass the ABO examination (Table 5.2). This

discrepancy between the two scoring methods indicates that the same cases that were scored by the plaster model would fail when using the new ABO software by the four examiners involved in the study (Table 5.3).

Plaster models

For the conventional plaster model method, the highest deduction was in the buccolingual inclination component among the examiners (mean ranging from 4.16 SD +/- 2.31 to 4.71 SD +/- 3.002). However, it was noted that the rest of the components did not score greatly lower than the buccolingual component except for the interproximal contact component, which had the lowest deduction with mean ranging from 0.226 SD +/-0.497 to 0.838 SD+/- 0.79 (Table 5.1).

In agreement with our findings, Okunami et al. (2007)and Costalos et al. (2005) reported that the plaster model interproximal contacts component showed the lowest points deducted.Costalos et al. (2005) reported that the buccolingual inclination was second to the alignment component for the highest deduction components. The authors explained that the examiners reported difficulties in consistently and accurately pinpointing the exact mesial and distal points to evaluate the alignment. On the contrary, this difficulty was not experienced in the current study.

Whereas Okunami et al. (2007) reported the occlusal relationship component as the highest deduction, it is important to note that the authors decided not to evaluate the buccolingual component in their study. It was noted that the difference in ranking of the components with the highest deduction among the above-mentioned studies (including the current study) was due to the minimal difference between the components deduction score. This may be explained by human variability in scoring.

Digital models

For the digital models software measured with the new software, the highest deduction was in the overjet and interproximal components (mean ranging from 25.903 SD +/- 3.037 to

24.080 SD +/- 4.229 and 27.871 SD +/- 4.514 to 6.548 SD +/- 4.780 respectively). This high deduction reported in the overjet component can be explained by the difficulty in indicating the occlusal surface centre point for the maxillary arch and the incisal edge point of the anterior teeth. In the mandibular arch it was also difficult to indicate the labial edge point. Whereas the high deduction reported in the interproximal component in the digital model was due to difficulty in identifying the contact with the adjacent teeth by the examiners. These difficulties reported are based on the verbal feedback given by the examiners in the study after using the new software system.

The lowest component deduction for the digital models was found to be in the buccolingual inclination component (mean ranging between 7.580 SD +/- 3.071 and 4.129 SD +/- 2.604). (Table 5.2). It is interesting to note that this component was found to have the highest component deducted in the plaster models (Table 5.1).

Okunami et al. (2007) and Costalos et al. (2005) reported that, similar to the plaster model, the interproximal contacts component showed the lowest points deducted for the digital models. This did not agree with the findings from the current study, indicating that the interproximal contacts component was found to have one of the highest deductions in the digital models software. This may be due to the difficulty reported by the examiners in identifying the contact point with the adjacent teeth.

In agreement with the current study, Costalos et al. (2005) found that the overjet component was among the highest deduction scores. The authors also reported that the occlusal contact component was the highest due to difficulties with the OrthoCAD software system that were reported to the manufacturer. Whereas Okunami et al. (2007) reported that the occlusal relationship and the alignment were the highest deduction scores for the digital models.

The descriptive data from the current study was not compared with Hildebrand et al.'s (2008) study, as the authors did not published the descriptive analysis.

It was noted that the mean total scores reported by Okunami et al.(2007) for both plaster (37.93 SD +/-11.02) and digital models (42.93 SD +/-9.56) were more than the maximum deduction pass suggested by the ABO (>30). In addition, Costalos et al. (2005) reported that the mean total score for plaster models (31.17 SD +/-10.47) was more than the maximum deduction pass suggested by the ABO while the digital models were in the upper limit of the grey zone (29.67 SD +/-9.29). The current study mean total ABO OGS score was low in the plaster models (ranging from 21.387 SD +/- 8.428 to 22.143 SD +/- 7.54) and high for the digital models (ranging from 89.323 SD +/- 14.549 to 124.097 SD +/- 12.918). This difference between the current study and the previously mentioned studies regarding the scores of the plaster models score (gold standard) may suggest that there is a variation in the treatment outcome of the models and the current study. On the other hand, the significantly higher ABO OGS score reported in the current study could be due to the use of the new software system in the current study.

6.2 Reliability

One of the aims of the current study was to assess the reliability of the ABO OGS scoring when measuring the seven components of the index using digital and study models. Inter-examiner and intra-examiner reliability was investigated.

6.2.1 Intra-examiner reliability

6.2.1.1 Plaster models

The appropriate training was undertaken by Examiner 1 for scoring ABO OGS plaster models using the calibrated ABO kit. Examiner 1 measured the 31 plaster models randomly twice on two separate occasions with a two-week interval between them. This gap was to ensure low potential for carry-over or recall effects (i.e. the first testing may influence the second). It was decided that a two-week interval would be a reasonable time span between repeated measurements.

High correlation was found for Examiner 1’s repeated measures (R = 0.915) using the

Pearson correlation coefficient test; this suggested high intra-examiner reliability for the total ABO OGS score using plaster models (Table 5.4). This correlation was found to be statistically significant (P = 0.00). In addition, statistically significant (P = 0.00) high intra-examiner reliability was found for all seven components of the ABO OGS, with the highest correlation reported for the (R = 0.982) and the lowest correlation coefficient reported for the interproximal contacts component (R = 0.771).

The intra-examiner correlation was found to be statistically significant, indicating that the correlation found was true and not due to chance. Therefore, it can be suggested from the current results that intra-examiner reliability was high for Examiner 1 when measuring the plaster models. It is important to realise that the intra-examiner reliability was assessed only for a single examiner; therefore, the results cannot be generalised.

6.2.1.2 Digital models

Moderate correlation was found for Examiner 1's repeated measures (R = 0.583) using the Pearson correlation coefficient test. This may suggest moderate intra-examiner reliability for the total ABO OGS score using digital models (Table 5.4). This correlation for the total new software score was found to be statistically significant (P = 0.0001). In addition, all seven components of the ABO OGS when assessed solely were found to have statistically significant correlation for the repeated measures from Examiner 1, except for the overjet component, which showed low correlation (R= 0.216).

The intra-examiner reliability was high for repeated measurements scored from plaster models in the current study. This can be explained by the appropriate training taken by the examiner using the ABO calibration models, as described in section 4.2.1. Although the intra-examiner reliability for the new software scores using digital models was statistically significant in most of the ABO components, it was relatively lower when compared with the plaster models' reliability.

6.2.1.3 Intra-examiner reliability comparison with previous studies

The intra-examiner reliability results of the current study agree with several published studies (Okunami et al., 2007, Hildebrand et al., 2008).Okunami et al. (2007) used the Wilcoxon test to assess the intra-examiner reliability for measurements taken for 10 digital and plaster models. The authors found no statistically significant difference between the repeated measurements for both plaster and digital. However, the authors reported that the plaster measurements' reliability was higher than the digital models, which was in agreement with the current study. It is worth mentioning that Okunami et al. (2007) eliminated the buccolingual inclination component of ABO OGS, because it showed errors in digital model assessment; this may have influenced the results.

Hildebrand et al. (2008) used the Spearman rank correlation coefficient to investigate the intra-examiner reliability for the ABO OGS scores for digital and plaster models. Intra-examiner analysis showed high reliability for both methods for each component of the ABO OGS and for the total score (R = 0.99). It is important to mention that only one examiner was involved in this study. This suggests that the authors depended solely on the intra-examiner repeatability to assess the reliability of the scores. The current study and Hildebrand et al. (2008) found high intra-examiner reliability for plaster models; however, in the current study a moderate correlation was found for the digital models. This can be explained by the use of a different software system in the current study, with which the examiners had limited training. Unlike the plaster models, no calibration was performed on the software digital models in the current study, which was not the case in Hildebrand et al.'s (2008) study, as they had used the OrthoCAD software package to calibrate the study examiner.

6.2.2 Inter-examiner reliability

Pearson correlation tests were used in the current study between different examiners in order to investigate the inter-examiner reliability between the four study examiners. The

four examiners had different levels of orthodontic background/knowledge; as such, the examiners did not consist only of orthodontists. This diverse examiner selection was made to investigate whether the new software scoring digital models can be used by non-orthodontists, which may help orthodontists to save time in clinic. This may also allow non-specialist researchers to participate in future research work that may involve using the digital model software to assess orthodontic treatment outcomes.

6.2.2.1 Plaster models

The appropriate training was taken by the four examiners for scoring ABO OGS plaster models using the calibrated ABO kit. All of the examiners measured the 31 plaster models.

The Pearson correlation coefficient, when used to analyse the measurements done on plaster models, showed a high inter-examiner reliability for the total ABO OGS score (Table 5.5). The highest correlation (ranging from R = 0.970 to 0.960 for the total score) was noted among Examiners 1, 2 and 3, who were all qualified dentists. However, slightly less correlation (ranging from R = 0.0877 to 0.814 for the total score) was noted between Examiner 4 (an undergraduate dental student) and the other 3 examiners. This may indicate that the level of dental and orthodontic knowledge may have a mild influence on the reliability of the ABO OGS plaster model scoring. However, the inter-examiner correlation was statistically significant (P = 0.000), suggesting high reliability among all examiners in scoring the plaster models.

It is worth mentioning that some components of the ABO OGS (e.g. inter-proximal contacts) had lower inter-examiner correlation when compared with the total ABO OGS score. The inter-examiner correlation was relatively lower in some components between Examiner 4 and the other examiners, reaching (R = 0.460). However, this correlation is still considered acceptable (moderate) and did not have a significant impact on the correlation of the total score.

6.2.2.2 Digital models

Pearson correlation tests were also used to analyse the inter-examiner reliability for the digital models. For this data set, we had very poor reliability among the four examiners (ranging from 0.415 to -0.469 for the total score), and the only moderate reliability was between Examiners 3 and Examiner 4 for the total score, with a correlation of 0.415 and significance of P=0.020, which is statistically significant (Table 5.6). Total score correlation was negative between some of the examiners, which indicates a statistically inversely correlated relation. In addition, the seven components of ABO OGS did not show a correlation among examiners.

These results from the current study may suggest high inter-examiner reliability for plaster models ABO OGS scoring among examiners and low inter-examiner reliability for the digital models in ABO OGS scoring. This can be explained by:

- No calibration being done for the digital models scoring, unlike the plaster models where the examiners were properly calibrated using the ABO training kit.
- Variation among the examiners in their skills in using computer advances and coping with it especially in terms of 3D viewing and pointing out landmarks.
- Dentists are more used to using plaster models through everyday work than digital models.

6.2.2.3 Inter-examiner reliability comparison with previous studies

The current study results did not agree with the Costalos et al. (2005) findings, which had used interclass correlation to test the inter-examiner reliability for ABO OGS scoring using plaster and digital models. They reported moderate reliability between examiners with slightly better reliability using the digital models (R = 0.69 and 0.65) when compared with plaster models (R = 0.53 and 0.46).

In the current study, failure to achieve a similar level of inter-examiner reliability for the

digital models was due to the use of new software, which the examiners did not calibrate. They simply used the new software by following the instructions by the manufacturer to indicate points. This lack of calibration for the ABO software may have influenced the results from the current study. Unlike the conventional methods no calibration was provided with the ABO software. In contrast, the examiners in the Costalos el al. (2005) study were trained using with the voice–over CD-ROM provided by the ABO, which was available in the OrthoCAD v.2.17 software package. In addition, calibration was performed for the two examiners involved in the Costalos et al. (2005) study by using two digital models and jointly reviewing their scoring after each analysis. However, Costalos et al. (2005)did not achieve the same high inter-examiner reliability for plaster models compared with the current study, due to the proper calibration and training undertaken in the current study using the ABO calibration kit for the plaster models.

However, Okunami et al. (2007) did not use statistical analysis to investigate inter-examiner reliability and instead measured five models solely by each examiner and then compared their results. When the discrepancy was found to be more than 2 points, the measurements were repeated. Although this technique ensured agreement between the examiners' measurements, it did not actually investigate the inter-examiner reliability of the plaster and digital software methods. It is worth mentioning that Hildebrand et al. (2008) did not report any inter-examiner reliability results, as only one examiner was involved in their study.

6.3 Validity of the new software system

The ABO software was assessed in the current study in order to investigate the level of agreement between the ABO OGS scores measured using the new software system and the conventional plaster models and evaluate if the new software can replace the conventional method. It was decided to consider the conventional method of using the plaster models and the ABO gauge as the gold standard for the following reasons:

- When the ABO developed the ABO OGS it was assessed for validity and reliability using the plaster model and the ABO gauge.
- The ABO board of examination has been accepting the ABO OGS on plaster models for the last 25 years until the present time.
- Several studies that were conducted in the literature have used the conventional method.

6.3.1 Statistical analysis used in comparing plaster and digital models

Several studies used different statistical analysis tests to assess the accuracy of digital models scoring compared to plaster models scoring (the gold standard). Interclass correlation and ANOVA were used in the Costalos et al. (2005) study, while Okunami et al. (2007) used the Wilcoxon rank-sum test for this comparison. Finally, Hildebrand et al. (2008) used Spearman correlation test, Wilcoxon rank-sum test and paired t-test.

It is important to highlight that good correlation does not in turn mean good agreement. Correlation analysis can be misleading, because correlation depends on the range of the measurements in the sample size. If the range is wide, the correlation will be greater than when the range is narrow. In the current study, it was decided to investigate the agreement between the conventional and digital methods using the Bland and Altman plots rather than the correlation tests. It was decided that the difference in the measurements between the digital software method and plaster models by more than 2 points per component was thought to be clinically significant. This magnitude was taken as a reference from a

previous study (Hildebrand et al.,2008).

6.3.2 Validity and accuracy of the new software

The new software used in the current study showed a lack of agreement with the plaster models for all components and the total score of the ABO OGS except for the buccolingual inclination component. This lack of agreement was demonstrated by the wide range of limits of agreement and means the difference of digital and plasters models in the Bland-Altman scatter plots.

It is obvious from the current results that the new software has low validity and cannot replace the conventional method of using plaster models, which is considered as the gold standard. The lack of agreement between digital methods and plaster models in the current study might be due to the following factors:

- Difficulty in localising the same landmarks on plaster and on the new software, which was also reported by Costalos et al. (2005).
- When measuring the ABO OGS components, the plaster is truly three-dimensional, while on a regular computer screen it is actually two-dimensional (Zilberman et al., 2003).
- To indicate the point on the software, occlusal views are automatically displayed during scoring. This may be changed when measuring models as the software has the option to zoom, rotate and pan. This means that indicating points was not always done with the same angle and can be rotated using different degrees of zooming. This may have introduced a source of distraction to an investigator who is not used to these 3D imaging facilities, thus creating a difference between digital and plaster models in indicating points.

Errors in the new software programming can lead to miscalculation of the ABO OGS scores. As such, there is a need for greater tolerance in the new software algorithms for

calculating the scores to allow for the microscopic nature of the digital model measurements

6.3.3 Comparison of current results with previous studies

Three studies were conducted previously to compare measurements from plaster and digital models for scoring ABO OGS (Table 2.3). Different versions of OrthoCAD software were used in these three studies.

Costalos et al. (2005) used a sample size of 24 models, and the ANOVA test demonstrated no statistically significant difference between the two methods in the total score and most of the components. The mean difference between the two methods was found to be 1.5 for the total ABO OGS score; this difference was found to be very small and was not clinically significant.

The current study findings did not agree with the results of Costalos et al. (2005). The results reported by Costalos et al. (2005) need to be interpreted with caution, as the authors calculated the mean difference between the two groups after adding the total scores for the total sample. This can be misleading, because the negative difference in a set of models could compensate for the positive difference in another set of models, which may lead to confusing results. In other words, the authors should have added the difference between the two methods for each set of models and then calculated the mean difference between the two methods.

Okunami et al. (2007) used the Wilcoxon rank-sum test to determine the statistical difference between the two methods for the 30 post-treatment models. The authors reported that the OrthoCAD program was not adequate to replace the conventional plaster models. The main reason behind the discrepancy in the software was found in the occlusal relationship and occlusal contacts components, which were found to be statistically significantly different from plaster models, which affected the study total score of ABO OGS. It is important to mention that the authors did not include the buccolingual component in the study.

Okunami et al. (2007) had difficulties with the maxillary and mandibular teeth because they overlap each other in occlusion. They attempted to solve this problem, but they found that some of the digital models had no contacts. They indicated that 3 to 35 points were deducted for the occlusal contacts, which caused a significant difference of the ABO OGS total score. For the occlusal relationship discrepancy the authors explained that this might have been due to measuring the plaster model in the perpendicular position, which is stated by the ABO guideline for measuring the occlusal relationship.

Moreover, Hildebrand et al. (2008) reported that the digital models version of ABO OGS cannot be a substitute for the conventional method. They indicated that alignment, occlusal contacts, overjet and the total score are statistically significantly different than plaster models. The authors demonstrated that digital models showed consistently higher scores than plaster ones. The greatest difference was found in the occlusal contacts component of the ABO OGS, which is 21 points. The digital models showed a 9 point difference.

In the current study, it was obvious that digital model software scoring showed a generally higher result. This is in agreement with Hildebrand et al. (2008), as the digital models displayed microscopic details, while the plaster displayed macroscopic details, which might be the reason for higher results in the digital software.

A systemic error also might be a cause of the lack of agreement between the two methods, as the digital model measurements were generally higher than plaster models, especially in occlusal contacts and overjet. Horton et al. (2010), when comparing tooth size between plaster and digital models, stated that the limitation of enlarging digital models might be a reason for the inaccuracy of plaster to digital models, since the new software had the option to zoom the model digitally. Scanning processes might be not accurate, and this also my lead to a lack of agreement between the methods, as Quimby et al. (2004) hypothesised in their study.

In summary, it seems that most of the studies except Costalos et al. (2005) agree that

currently the digital models cannot replace the conventional plaster method, mainly due to technical issues with the different software systems used. It is interesting to note that the four studies (including the current study) did not agree on which ABO OGS components might be causing the greatest discrepancy between the two methods.

6.4 Time for ABO OGS scoring

The current study was the first to compare between the digital and plaster models with regards to the time taken for scoring the ABO OGS. The comparison was performed to investigate which method is faster and if there is a significant difference between the two methods. In addition, a comparison between the times taken for scoring by different examiners involved in the study was performed.

In the current study, the 4 examiners' scoring times (each scored 31 models) were added together to represent a total of (31x4) 124 scoring attempts by the four examiners (Table 5.11). There was a statistically significant (P = 0.000) increase in the time taken for scoring digital models software compared with plaster models. The mean time of measuring plaster models was 13.71 minutes compared with 34.14 minutes for the digital models software, with a mean difference of 20.43 minutes. This difference is considered to be clinically significant.

Different studies were designed to compare between plaster and digital models for the time taken to score different types of indices, including Bolton ratio analysis and Peer Assessment Exercise (PAR). The current study findings agree with Mayers et al. (2005), who reported a statistically significant increase in the time taken for scoring using digital models software compared with plaster models. However, Mayers et al. (2005) investigated the PAR index rather than the ABO OGS as an index for assessing the treatment outcome.

In contrast, Mullen et al. (2007) reported a decrease in the time taken to measure the Bolton ratio using digital models software when compared with plaster models. However, the difference in time was only 65 seconds. Horton et al. (2010) compared the time taken to

measure mesial and distal tooth width using plaster and digital models, and they indicated that the digital models were faster than plaster models.

There may be several reasons that explain the extended time for ABO OGS scoring using the digital model software, including:

- The software was not user-friendly; a minor error during scoring may necessitate starting all over again.
- The examiners did not have enough experience in handling digital models when compared with plaster models. This may require more time for scoring.
- Localising an anatomic point on a tooth surface may require several manoeuvres and zooming using digital models, which may be time-consuming, especially for inexperienced examiners.

  On the other hand, it is important to note that requesting and refilling plaster models may also be time-consuming, while digital models are easier to access (Mayers et al., 2005).

6.4.1 Comparison among examiners for scoring time

6.4.1.1 Time scoring plaster models

For plaster models, the results showed no statistically significant difference among examiners. However, Examiners 1, 2 and 3 had mean times for scoring of less than 12 minutes, while Examiner 4 had a mean of 20.41 minutes for scoring the plaster models (Table 5.10). Although this difference in time between Examiner 4 and the other examiners was not found to be statistically significant, it is believed to be clinically significant (almost double). Examiner 4 (dental undergraduate student) had the least background in the dentistry field compared with the other examiners; this may be an influencing factor for the difference in scoring time.

6.4.1.2 Time scoring digital models

No statistically significant difference was found among examiners for the scoring time using digital models software. Examiner 1 had the lowest mean time for scoring (30.94 minutes) when compared with the other examiners (ranging from 36.03 to 30.94 minutes) (Table 5.9). This may be due to more training being performed by Examiner 1, who repeated the measurement twice in the intra-examiner reliability analysis. However, this minor difference was not found to be statistically significant.

These findings may suggest that a variety of investigators with different backgrounds in dentistry can use the ABO OGS software with acceptable clinical and statistical scoring outcomes. However, more research in this area may be required to confirm our findings.

6.5 Sources of error in the ABO software system

The main source of error in this study was almost certainly due to landmark identification. It was difficult to select the same points of measurements on the model each time, and this problem was noticeable in both upper and lower arch measurements. This has been found in other studies (Mullen et al., 2007, Redlich et al., 2008). The repeated landmarks location using either method can differ, and the examiner opinion of the precise location of a point may vary at random. Even with the help of the rotation function and magnification in the software, accurate point location remains difficult.

6.6 Clinical implications

The results from the current study suggest that the new software system is not clinically acceptable as a substitute for the conventional plaster models method. That is because the ABO software system lacks validity and reliability and this will have a significant impact on the evaluation of treatment outcome, especially for the ABO examination and research work. There is a significant chance that candidates submitting their cases to the ABO examination board using the conventional plaster models and scored as having passed will fail if they scored the same cases using the new software scoring system.

It is obvious form the data represented in Table 5.8 that less than one third of the sample were given a fail score using the conventional plaster method, while all the sample (100%) failed to pass using ABO software . This data may suggest that using the ABO software for the ABO examination will result in higher failure rate with at least triple the failure reported using the conventional plaster method. This clinically significant difference in the pass and fail rate in this study between the two scoring methods was confirmed statistically using Chi square test where the difference was found to be highly significant.

In addition, there seems to be no benefit from the use of the digital models to score the ABO OGS with regards to saving time for the clinician and the researcher. This may cause a debate on the cost effectiveness of the software system.

Chapter 7: Conclusion

It can be concluded from the current study results that:

- There was very low agreement between the new software and the conventional plaster models in all of the ABO OGS components except for the buccolingual component. Therefore, the new software digital model software is not a valid method for ABO OGS scoring; it cannot replace the conventional plaster models method.
- The new software has acceptable intra-examiner reliability; however, inter-examiner reliability was found to be low.
- The conventional plaster models ABO OGS scoring has high intra-examiner and inter-examiner reliability.
- There is a low level of agreement in ABO OGS using the new software and digital models between orthodontists and non-orthodontists.
- There is a high level of agreement in ABO OGS using the conventional plaster models between orthodontists and non-orthodontists.
- The new software requires significantly more time for ABO OGS scoring than the conventional plaster models method.
- The level of dental background does not appear to have a significant influence on the time needed to use the new software.
- Undergraduate dental students may require a longer time than postgraduate-level dentists or orthodontists to score the ABO OGS using the conventional plaster models method. This difference was found to be clinically significant.

References

ABIZADEH, N., MOLES, D. R., O'NEILL, J. & NOAR, J. H. 2012. Digital versus plaster study models: How accurate and reproducible are they? *Journal of orthodontics,* 39**,** 151-159.

ACKERMAN, J. L. & PROFFIT, W. R. 1969. Characteristics of malocclusion- A modern approach to classification and diagnosis. *American Journal of Orthodontics,* 56**,** 443-&.

ALLEN, M. J. & YEN, W. M. 2001. *Introduction to measurement theory*, Waveland Press.

ANGLE, E. H. 1899. Classification of malocclusion. *Dent Cosmos,* 41**,** 248-64.

BALLARD, C. F. & WAYMAN, J. B. 1965. A report on a survey of the orthodontic requirements of 310 army apprentices. *The Dental practitioner and dental record,* 15**,** 221-6.

BELL, A., AYOUB, A. F. & SIEBERT, P. 2003. Assessment of the accuracy of a three-dimensional imaging system for archiving dental study models. *Journal of Orthodontics,* 30**,** 219-223.

BELL, S. 2001. *A beginner's guide to uncertainty of measurement*, National Physical Laboratory Teddington, Middlesex.

BJÖRK, A., KREBS, A. & SOLOW, B. 1964. A Method for Epidemiological Registration of Malocculusion. *Acta Odontologica Scandinavica,* 22**,** 27-41.

BLAND, J. M. & ALTMAN, D. G. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet,* 1**,** 307-310.

BLAND, J. M. & ALTMAN, D. G. 1999. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research,* 8**,** 135-160.

BORZABADI-FARAHANI, A. & BORZABADI-FARAHANI, A. 2011. Agreement between the index of complexity, outcome, and need and the dental and aesthetic

components of the index of orthodontic treatment need. *American Journal of Orthodontics and Dentofacial Orthopedics,* 140**,** 233-238.

BOSSUYT, P. M., REITSMA, J. B., BRUNS, D. E., GATSONIS, C. A., GLASZIOU, P. P., IRWIG, L. M., MOHER, D., RENNIE, D., DE VET, H. C. W. & LIJMER, J. G. 2003. The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration. *Annals of Internal Medicine,* 138**,** W1-12.

BROOK, P. H. & SHAW, W. C. 1989. The development of an index of orthodontic treatment priority. *European Journal of Orthodontics,* 11**,** 309-320.

CARLOS, J. P. 1970. Evaluation of indices of malocclusion. *International Dental Journal,* 20**,** 606-&.

CASKO, J. S., VADEN, J. L., KOKICH, V. G., DAMONE, J., JAMES, R. D., CANGIALOSI, T. J., RIOLO, M. L., OWENS, S. E. & BILLS, E. D. 1998. Objective grading system for dental casts and panoramic radiographs. *American Journal of Orthodontics and Dentofacial Orthopedics,* 114**,** 589-599.

CONS, N. C., JENNY, J. & KOHOUT, F. J. 1986. *DAI--the Dental Aesthetic Index,* College of Dentistry, University of Iowa Iowa City.

COSTALOS, P. A., SARRAF, K., CANGIALOSI, T. J. & EFSTRATIADIS, S. 2005. Evaluation of the accuracy of digital model analysis for the American Board of Orthodontics objective grading system for dental casts. *American Journal of Orthodontics and Dentofacial Orthopedics,* 128**,** 624-629.

CREED, B., KAU, C. H., ENGLISH, J. D., XIA, J. J. & LEE, R. P. 2011. A Comparison of the Accuracy of Linear Measurements Obtained from Cone Beam Computerized Tomography Images and Digital Models. *Seminars in Orthodontics,* 17**,** 49-56.

DANIELS, C. & RICHMOND, S. 2000. The development of the index of complexity, outcome and need (ICON). *Journal of orthodontics,* 27**,** 149-62.

DRAKER, H. 1967. American association of orthodontists approval of the assessment record form and the definition of handicapping malocclusion. *Journal of American Dental Association,* 75**,** 1441-1442.

DRAKER, H. L. 1960. Handicapping labio-lingual deviations: a proposed index for public health purposes. *American Journal of Orthodontics,* 46**,** 295-305.

FIRESTONE, A. R., BECK, F. M., BEGLIN, F. M. & VIG, K. W. L. 2002. Evaluation of the peer assessment rating (PAR) index as an index of orthodontic treatment need. *American Journal of Orthodontics and Dentofacial Orthopedics,* 122**,** 463-469.

FLEMING, P. S., MARINHO, V. & JOHAL, A. 2011. Orthodontic measurements on digital study models compared with plaster models: a systematic review. *Orthodontics & Craniofacial Research,* 14**,** 1-16.

GOONEWARDENE, R. W., GOONEWARDENE, M. S., RAZZA, J. M. & MURRAY, K. 2008. Accuracy and validity of space analysis and irregularity index measurements using digital models. *Australian Orthodontic Journal,* 24**,** 83-90.

HAJEER, M. Y., MILLETT, D. T., AYOUB, A. F. & SIEBERT, J. P. 2004. Applications of 3D imaging in orthodontics: part II. *Journal of orthodontics,* 31**,** 154-62.

HAMDAN, A. M. & ROCK, W. P. 1999. An appraisal of the Peer Assessment Rating (PAR) Index and a suggested new weighting system. *European Journal of Orthodontics,* 21**,** 181-192.

HANNEMAN, S. K. 2008. Design, analysis, and interpretation of method-comparison studies. *AACN advanced critical care,* 19**,** 223-34.

HILDEBRAND, J. C., PALOMO, J. M., PALOMO, L., SIVIK, M. & HANS, M. 2008. Evaluation of a software program for applying the American Board of Orthodontics objective grading system to digital casts. *American Journal of Orthodontics and Dentofacial Orthopedics,* 133**,** 283-289.

HORTON, H. M. I., MILLER, J. R., GAILLARD, P. R. & LARSON, B. E. 2010. Technique Comparison for Efficient Orthodontic Tooth Measurements Using Digital Models. *Angle Orthodontist,* 80**,** 254-261.

JAMES, R. D. 2002. Objective cast and panoramic radiograph grading system. *American Journal of Orthodontics and Dentofacial Orthopedics,* 122**,** 450-450.

KEATING, A. P., KNOX, J., BIBB, R. & ZHUROV, A. I. 2008. A comparison of plaster, digital and reconstructed study model accuracy. *Journal of Orthodontics,* 35**,** 191-201.

LEIFERT, M. F., LEIFERT, M. M., EFSTRATIADIS, S. S. & CANGIALOSI, T. J. 2009. Comparison of space analysis evaluations with digital models and plaster dental casts. *American Journal of Orthodontics and Dentofacial Orthopedics,* 136.

LINDER-ARONSON, S. 1974. Orthodontics in the Swedish Public Dental Health Service. *Transactions. European Orthodontic Society***,** 233-40.

LITTLE, R. M. 1975. Irregularity index-Quantitative score of mandibular anterior alignment. *American Journal of Orthodontics and Dentofacial Orthopedics,* 68**,** 554-563.

LLEWELLYN, S. K., HAMDAN, A. M. & ROCK, W. P. 2007. An index of orthodontic treatment complexity. *European Journal of Orthodontics,* 29**,** 186-192.

LUU, N. S., NIKOLCHEVA, L. G., RETROUVEY, J.-M., FLORES-MIR, C., EL-BIALY, T., CAREY, J. P. & MAJOR, P. W. 2012. Linear measurements using virtual study models: A systematic review. *The Angle Orthodontist,* 82**,** 1098-1106.

MAYERS, M., FIRESTONE, A. R., RASHID, R. & VIG, K. W. L. 2005. Comparison of peer assessment rating (PAR) index scores of plaster and computer-based digital models. *American Journal of Orthodontics and Dentofacial Orthopedics,* 128**,** 431-434.

MCGUINNESS, N. J. & STEPHENS, C. D. 1992. Storage of orthodontic study models in hospital units in the U.K. *Journal of Orthodontics,* 19**,** 227-32.

MESSICK, S. 1989. Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher,* 18**,** 5-11.

MOL, B. W., LIJMER, J. G., EVERS, J. L. H. & BOSSUYT, P. M. M. 2003. Characteristics of good diagnostic studies. *Seminars in Reproductive Medicine,* 21**,** 17-25.

MULLEN, S. R., MARTIN, C. A., NGAN, P. & GLADWIN, M. 2007. Accuracy of space analysis with emodels and plaster models. *American Journal of Orthodontics and Dentofacial Orthopedics,* 132**,** 346-352.

OKUNAMI, T. R., KUSNOTO, B., BEGOLE, E., EVANS, C. A., SADOWSKY, C. & FADAVI, S. 2007. Assessing the American Board of Orthodontics objective grading system: Digital vs plaster dental casts. *American Journal of Orthodontics and Dentofacial Orthopedics,* 131**,** 51-56.

ONYEASO, C. O. & BEGOLE, E. A. 2007. Relationship between index of complexity, outcome and need, dental aesthetic index, peer assessment rating index, and American Board of Orthodontics objective grading system. *American Journal of Orthodontics and Dentofacial Orthopedics,* 131**,** 248-252.

PEAT, J. 2002. Health services research: a handbook of quantitative methods. London, Sage.

QUIMBY, M. L., VIG, K. W. L., RASHID, R. G. & FIRESTONE, A. R. 2004. The accuracy and reliability of measurements made on computer-based digital models. *Angle Orthodontist,* 74**,** 298-303.

REDLICH, M., WEINSTOCK, T., ABED, Y., SCHNEOR, R., HOLDSTEIN, Y. & FISCHER, A. 2008. A new system for scanning, measuring and analyzing dental casts based on a 3D holographic sensor. *Orthodontics & Craniofacial Research,* 11**,** 90-95.

RICHMOND, S., SHAW, W. C., O'BRIEN, K. D., BUCHANAN, I. B., JONES, R., STEPHENS, C. D., ROBERTS, C. T. & ANDREWS, M. 1992a. The development of the PAR index peer assessment rating reliability and validity. *European Journal of Orthodontics,* 14**,** 125-130.

RICHMOND, S., SHAW, W. C., ROBERTS, C. T. & ANDREWS, M. 1992b. The PAR Index (Peer Assessment Rating): methods to determine outcome of orthodontic treatment in terms of improvement and standards. *European journal of orthodontics,* 14**,** 180-7.

SANTORO, M., GALKIN, S., TEREDESAI, M., NICOLAY, O. F. & CANGIALOSI, T. J. 2003. Comparison of measurements made on digital and plaster models. *American Journal of Orthodontics and Dentofacial Orthopedics*, 124**,** 101-105.

SHAW, W. C., RICHMOND, S. & OBRIEN, K. D. 1995. The use of occlusal indexes- A european perspective. *American Journal of Orthodontics and Dentofacial Orthopedics,* 107**,** 1-10.

SOUSA, M. V. S., VASCONCELOS, E. C., JANSON, G., GARIB, D. & PINZAN, A. 2012. Accuracy and reproducibility of 3-dimensional digital model measurements. *American Journal of Orthodontics and Dentofacial Orthopedics,* 142**,** 269-273.

STEVENS, D. R., FLORES-MIR, C., NEBBE, B., RABOUD, D. W., HEO, G. & MAJOR, P. W. 2006. Validity, reliability, and reproducibility of plaster vs digital study models: Comparison of peer assessment rating and Bolton analysis and their constituent measurements. *American Journal of Orthodontics and Dentofacial Orthopedics,* 129**,** 794-803.

SUMMERS, C. J. 1971. Occlusal index - System for identifying and scoring occlusal disorders. *American Journal of Orthodontics,* 59**,** 552-&.

TOMASSETTI, J. J., TALOUMIS, L. J., DENNY, J. M. & FISCHER, J. R. 2001. A comparison of 3 computerized Bolton tooth-size analyses with a commonly used method. *Angle Orthodontist,* 71**,** 351-357.

VAN DER MEER, W. J., ANDRIESSEN, F. S., WISMEIJER, D. & REN, Y. 2012. Application of intra-oral dental scanners in the digital workflow of implantology. *PLoS ONE,* 7**,** e43312.

VEENEMA, A., KATSAROS, C., BOXUM, S., BRONKHORST, E. & KUIJPERS-JAGTMAN, A. 2009. Index of Complexity, Outcome and Need scored on plaster and digital models. *The European Journal of Orthodontics,* 31**,** 281-286.

WATANABE-KANNO, G. A., ABRÃO, J., MIASIRO JUNIOR, H., SÁNCHEZ-AYALA, A. & LAGRAVÈRE, M. O. 2009. Reproducibility, reliability and validity of measurements obtained from Cecile3 digital models. *Brazilian Oral Research,* 23**,** 288-295.

WHITING, P., RUTJES, A. W. S., REITSMA, J. B., GLAS, A. S., BOSSUYT, P. M. M. & KLEIJNEN, J. 2004. Sources of Variation and Bias in Studies of Diagnostic AccuracyA Systematic Review. *Annals of Internal Medicine,* 140**,** 189-202.

WHITING, P. F., RUTJES, A. W. S., WESTWOOD, M. E., MALLETT, S. & GRP, Q.-S. 2013. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *Journal of Clinical Epidemiology,* 66**,** 1093-1104.

YOUNIS, J. W., VIG, K. W. L., RINCHUSE, D. J. & WEYANT, R. J. 1997. A validation study of three indexes of orthodontic treatment need in the United States. *Community Dentistry and Oral Epidemiology,* 25**,** 358-362.

ZILBERMAN, O., HUGGARE, J. A. V. & PARIKAKIS, K. A. 2003. Evaluation of the validity of tooth size and arch width measurements using conventional and three-dimensional virtual orthodontic models. *Angle Orthodontist,* 73**,** 301-306.

Appendix 1: Bland-Altman plots and tables for plaster and digital models

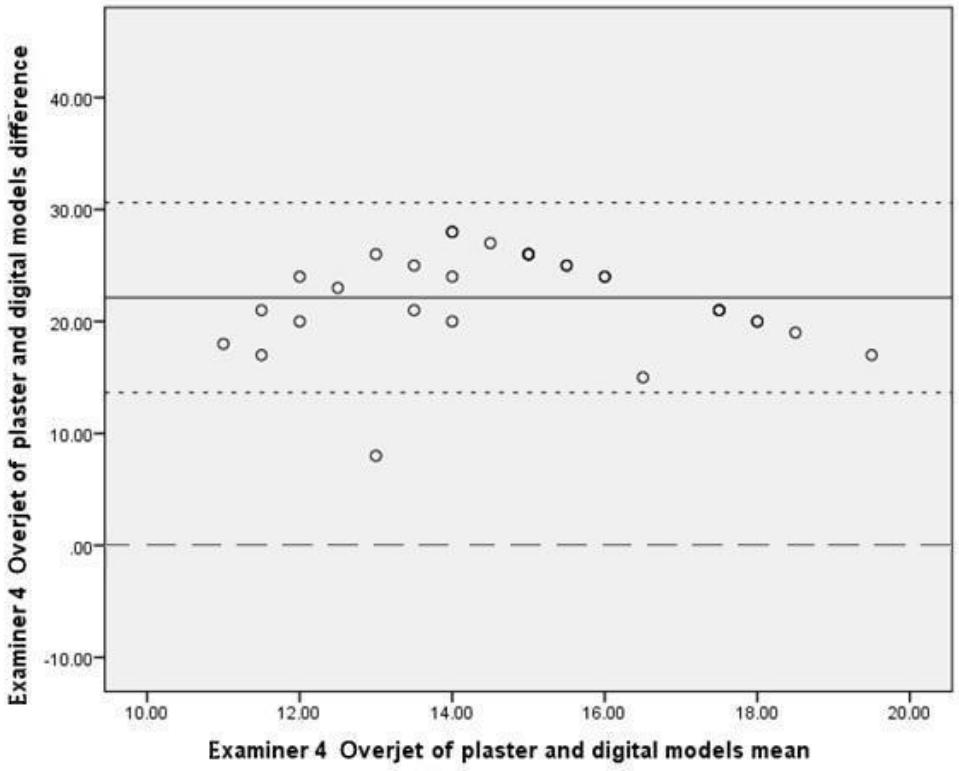


Figure 5.15 Bland-Altman scatter plot for Alignment for Examiner 2

Table 5.12 Mean and limits of agreement of plaster and digital models for Alignment for Examiner 2

The mean difference for the Alignment for Examiner 2	4.32
The Limits of agreements of Alignment for Examiner 2	13.45-(-4.80)

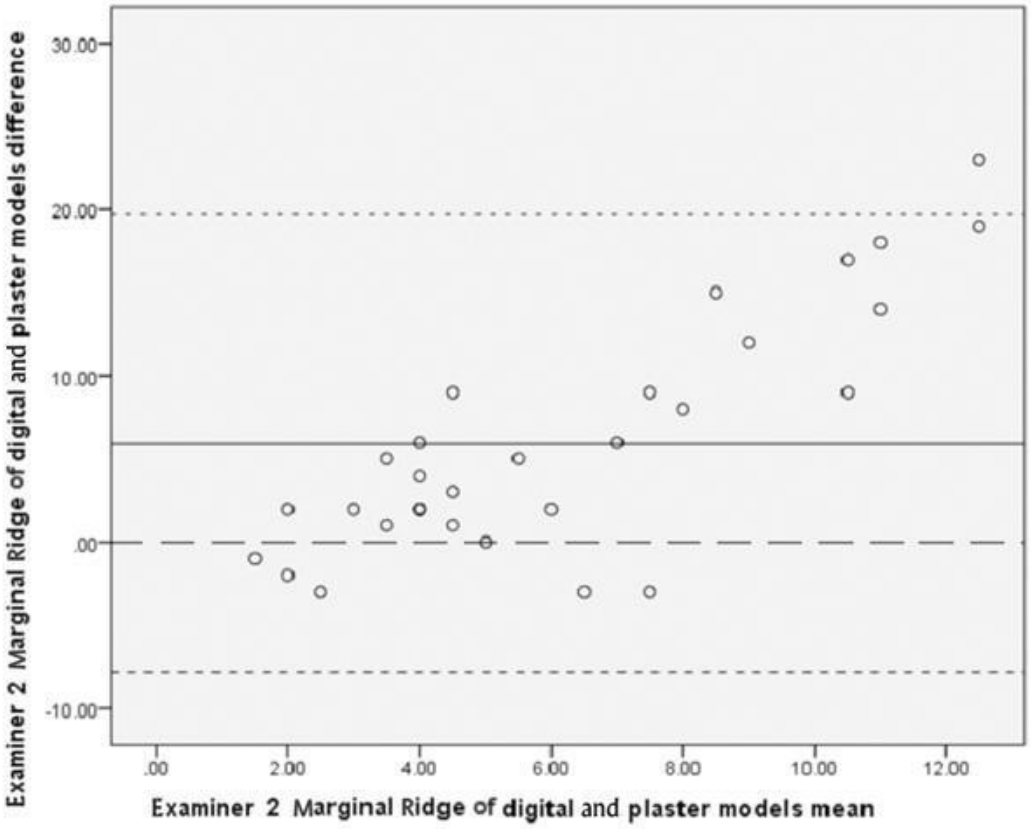


Figure 5.16 Bland-Altman scatter plot for Marginal Ridges for Examiner 2

Table 5.13 Mean and limits of agreement of plaster and digital models for Marginal Ridge for Examiner 2

The mean difference of the Marginal Ridge for Examiner 2	5.93
The limits of agreements of Alignment for Examiner 2	19.73-(-7.86)

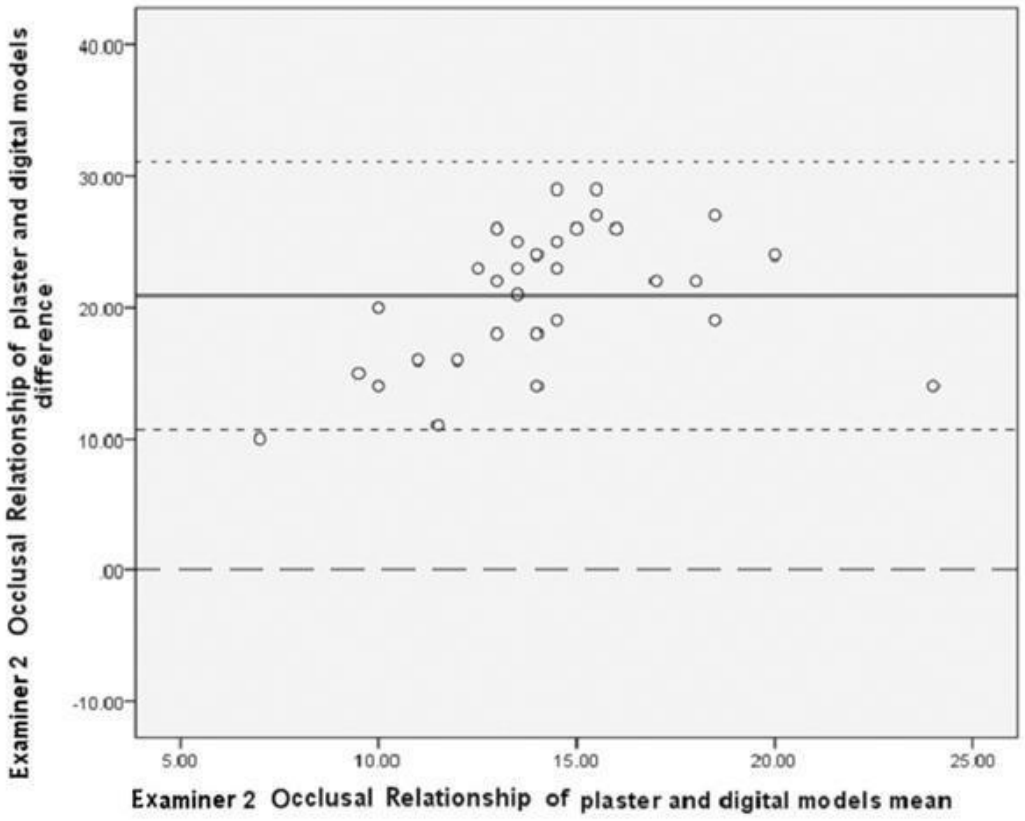


Figure 5.17 Bland-Altman scatter plot for Occlusal Relationship for Examiner 2

Table 5.14 Mean and limits of agreement of plaster and digital models for Occlusal Relationship for Examiner 2

The mean difference of the Occlusal Relationship for Examiner 2	20.90
The limits of agreements of Occlusal Relationship for Examiner 2	31.15 –(10.65)

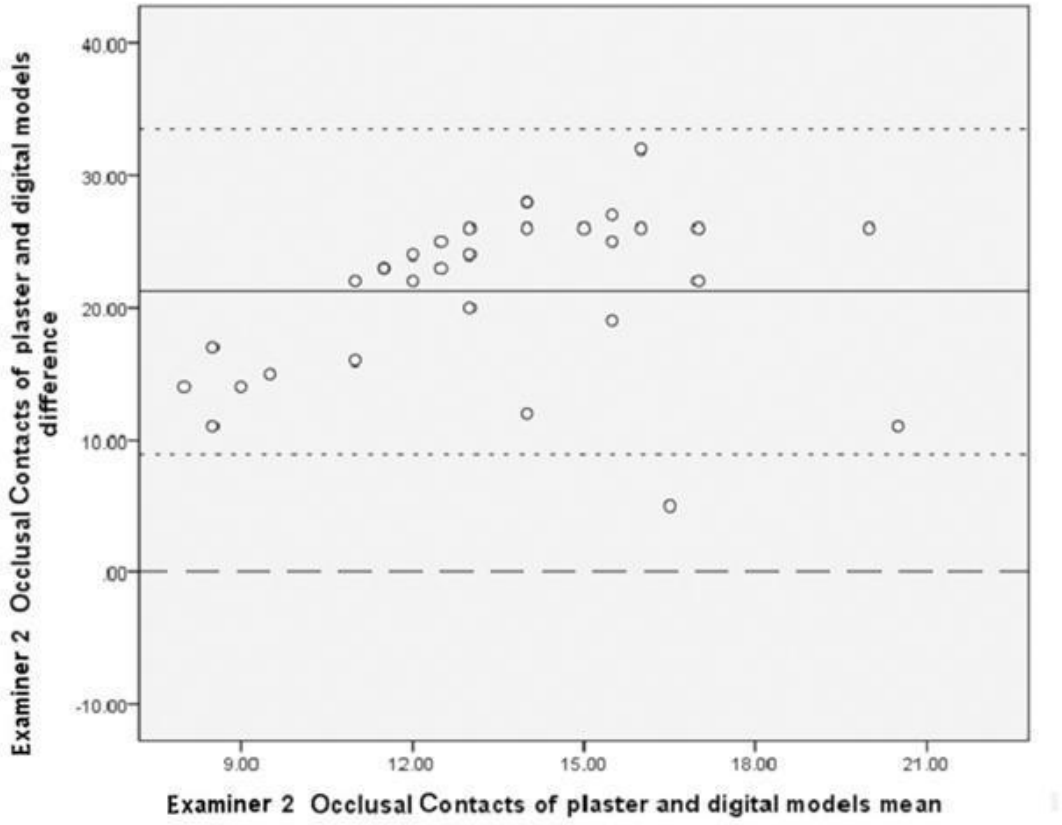


Figure 5.18 Bland-Altman scatter plot for Occlusal Contacts for Examiner 2

Table 5.15 Mean and limits of agreement of plaster and digital models for Occlusal Contacts for Examiner 2

The mean difference of the Occlusal Contacts for Examiner 2	21.22
The limits of agreements of Occlusal Contacts for Examiner 2	33.54-(8.9)

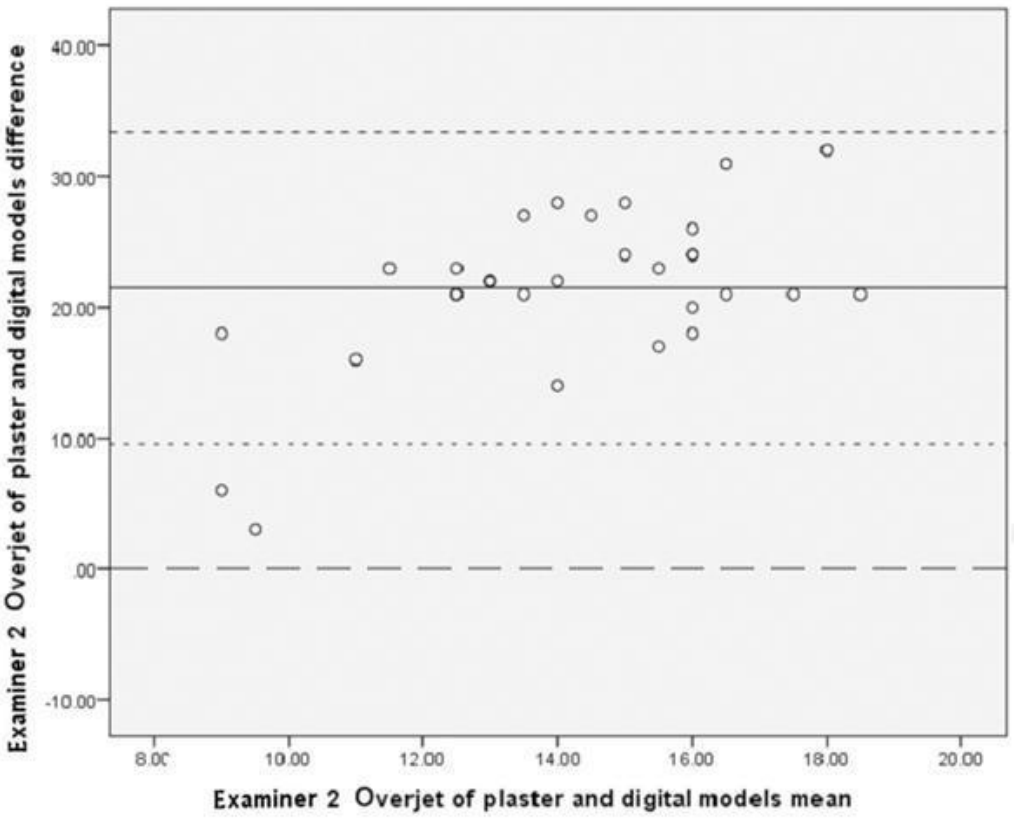


Figure 5.19 Bland-Altman scatter plot for Overjet for Examiner 2

Table 5.16 Mean and limits of agreement of plaster and digital models for Overjet for Examiner 2

The mean difference of the Overjet for Examiner 2	21.48
The limits of agreements of Overjet for Examiner 2	33.4-(9.56)

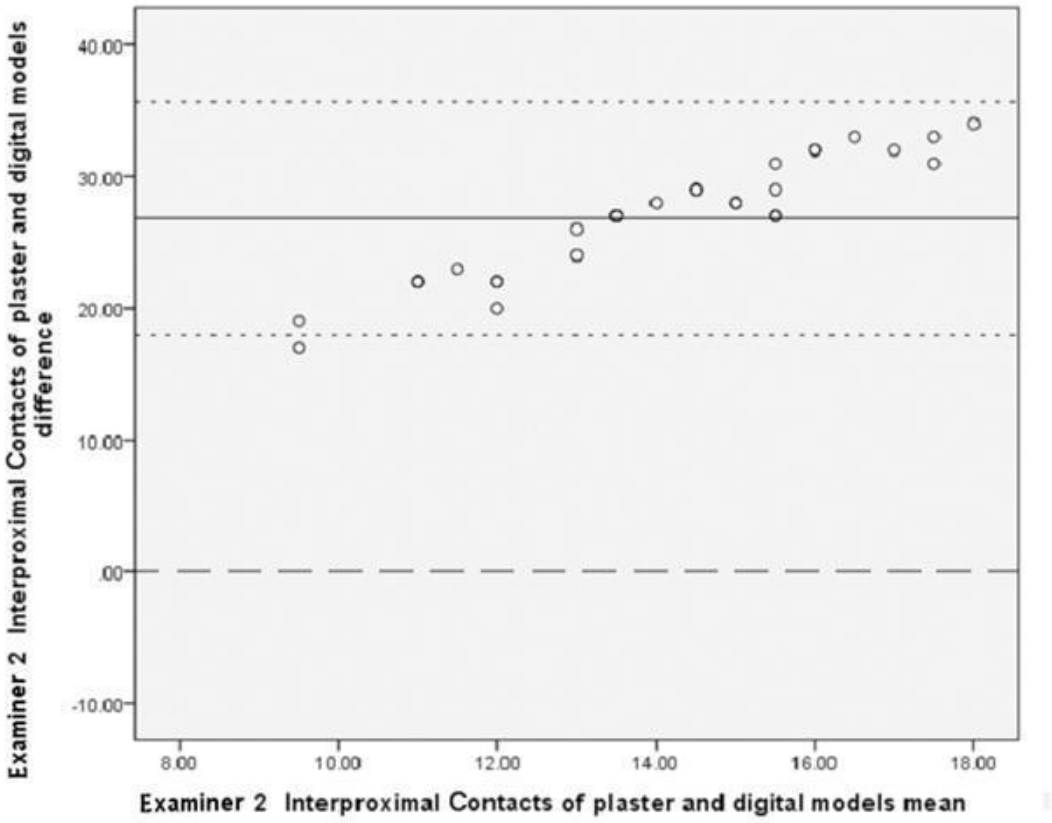


Figure 5.20 Bland-Altman scatter plot for Interproximal Contacts for Examiner 2

Table 5.17 Mean and limits of agreement of plaster and digital models for Interproximal Contacts for Examiner 2

The mean difference of the Interproximal Contacts for Examiner 2	26.83
The limits of agreements of Interproximal Contacts Examiner 2	33.66-(18.00)

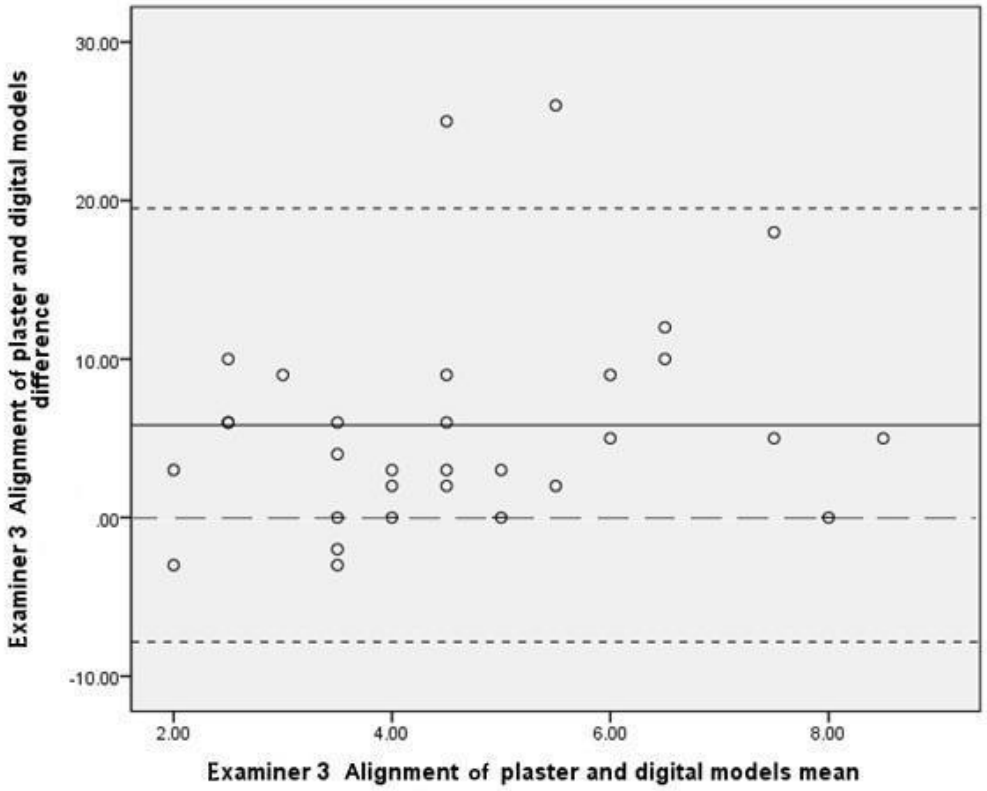


Figure 5.21 Bland-Altman scatter plot for Alignment for Examiner 3

Table 5.18 Mean and limits of agreement of plaster and digital models for Alignment for Examiner 3

The mean difference of the Alignment for Examiner 3	8.83
The limits of agreements of Alignment for Examiner 3	19.51– (-7.83)

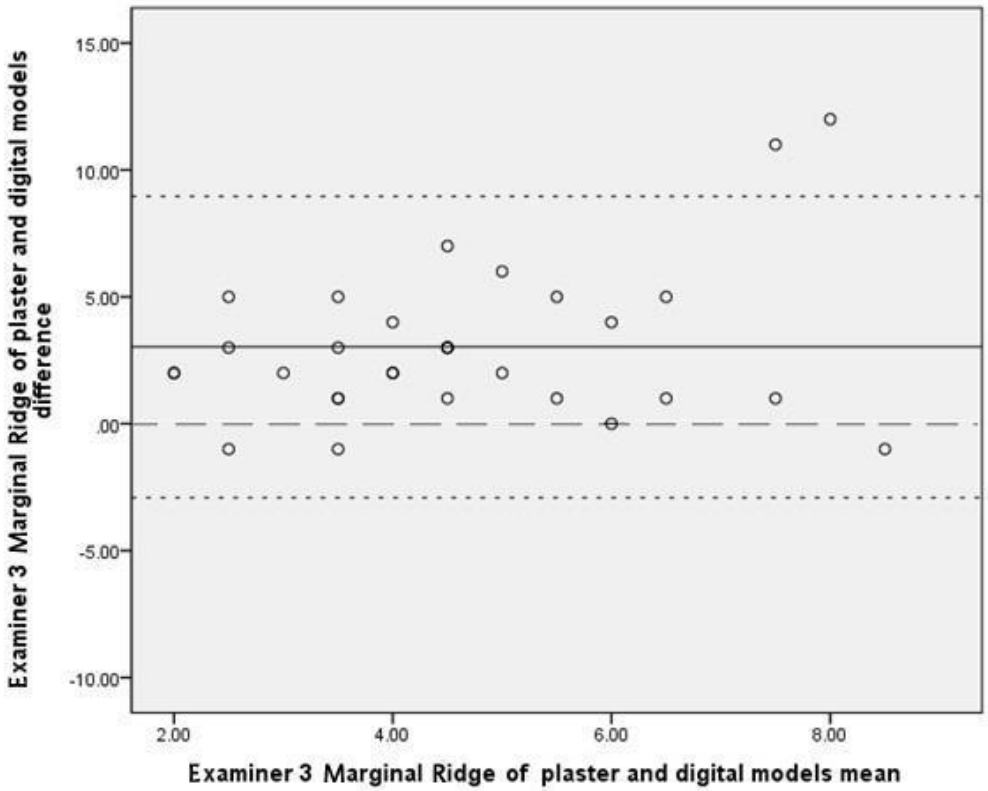


Figure 5.22 Bland-Altman scatter plot for Marginal Ridges for Examiner 3

Table 5.19 Mean and limits of agreement of plaster and digital models for Marginal Ridge for Examiner 3

The mean difference of the Marginal Ridge for Examiner 3	3.03
The limits of agreements of Marginal Ridges for Examiner 3	8.86-(-2.90)

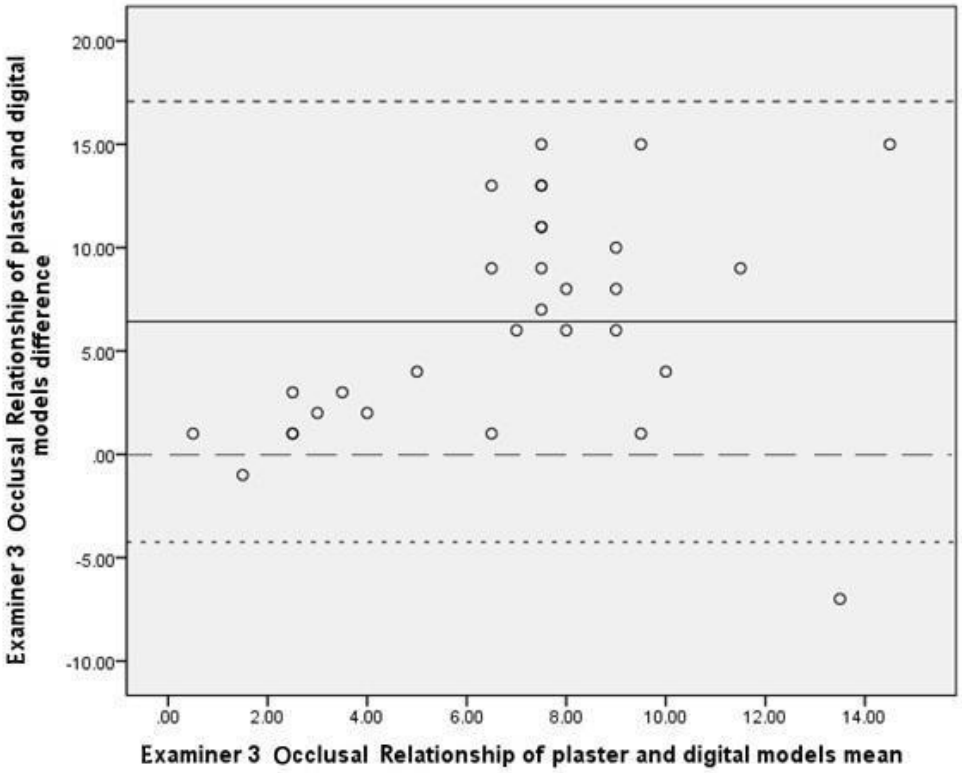


Figure 5.23 Bland-Altman scatter plot for Occlusal Relationship of Examiner 3

Table 5.20 Mean and limits of agreement of plaster and digital models for Occlusal Relationship for Examiner 3

The mean difference of the Occlusal Relationship for Examiner 3	6.41
The limits of agreements of Occlusal Relationship for Examiner 3	17.07-(-4.2)

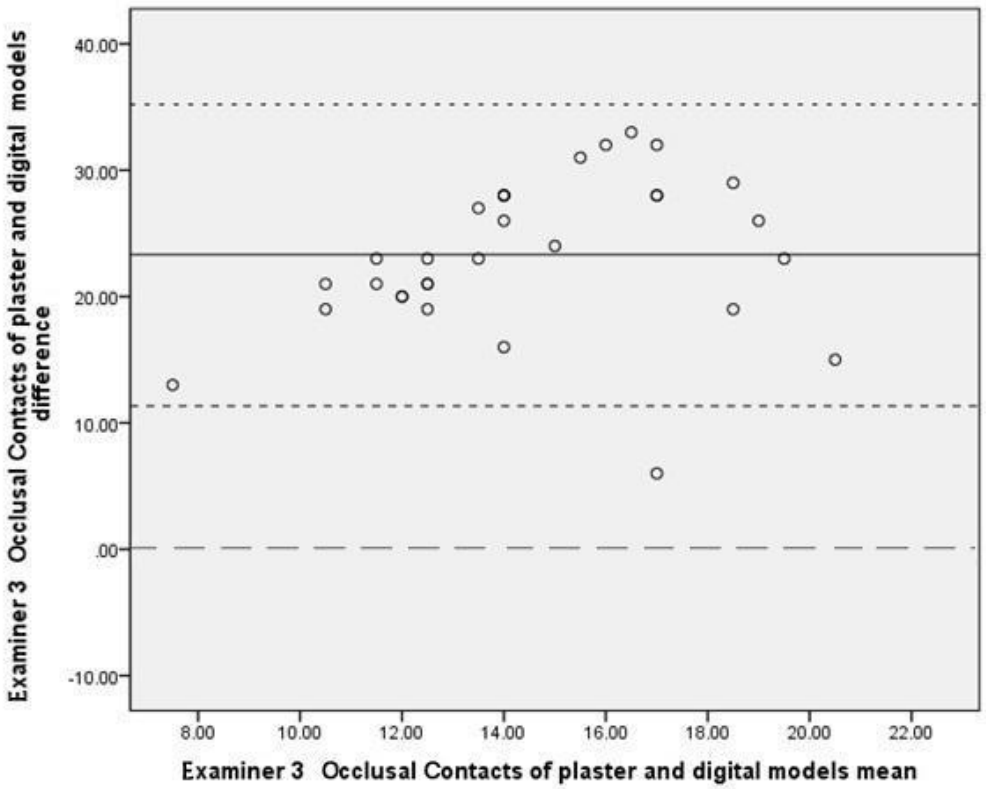


Figure 5.24 Bland-Altman scatter plot for Occlusal Contact for Examiner 3

Table 5.21 Mean and limits of agreement of plaster and digital models for Occlusal Contacts for Examiner 3

The mean difference of the Occlusal Contacts for Examiner 3	23.32
The limits of agreements of Occlusal Contacts for Examiner 3	35.20 – 11.43

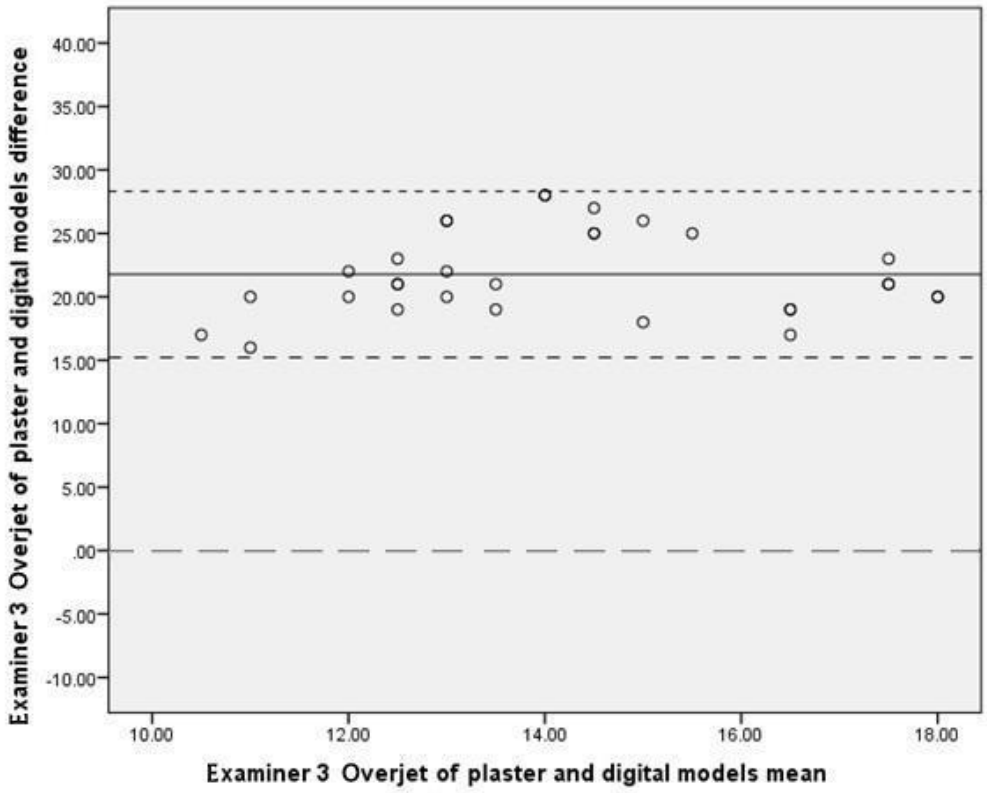


Figure 5.25 Bland-Altman scatter plot for Overjet for Examiner 3

Table 5.22 Mean and limits of agreement of plaster and digital models for Overjet for Examiner 3

The mean difference of the Overjet for Examiner 3	21.77
The limits of agreements of Overjet for Examiner 3	28.32 -15.22

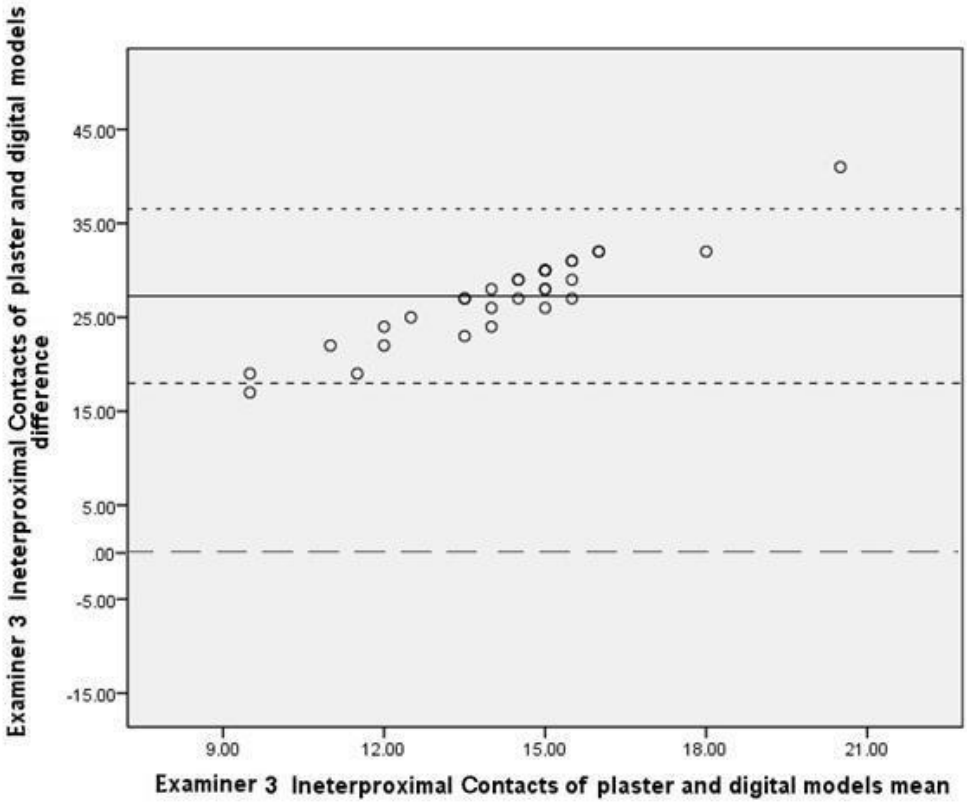


Figure 5.26 Bland-Altman scatter plot for Interproximal Contacts for Examiner 3

Table 5.23 Mean and limits of agreement of plaster and digital models for Interproximal Contacts for Examiner 3

The mean difference of the Interproximal Contacts for Examiner 3	27.25
The limits of agreements of the Interproximal Contacts for Examiner 3	35.51-17.99

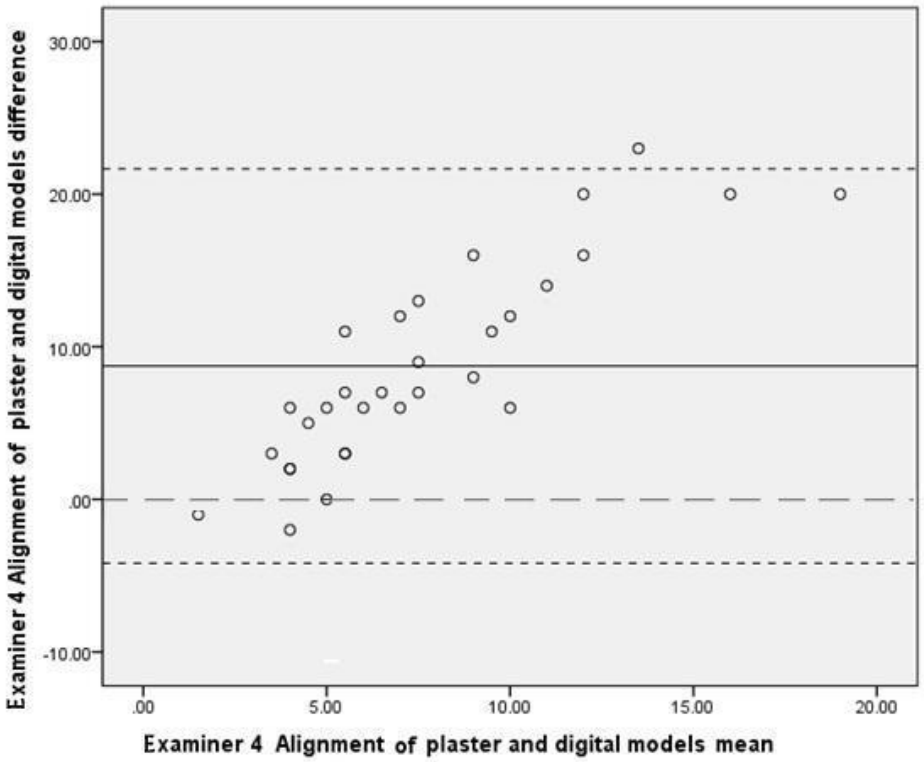


Figure 5.27 Bland-Altman scatter plot for Alignment for Examiner 4

Table 5.24 Mean and limits of agreement of plaster and digital models for Alignment for Examiner 4

The mean difference of the Alignment for Examiner 4	8.74
The limits of agreements of alignment Examiner 4	21.66-(-4.18)

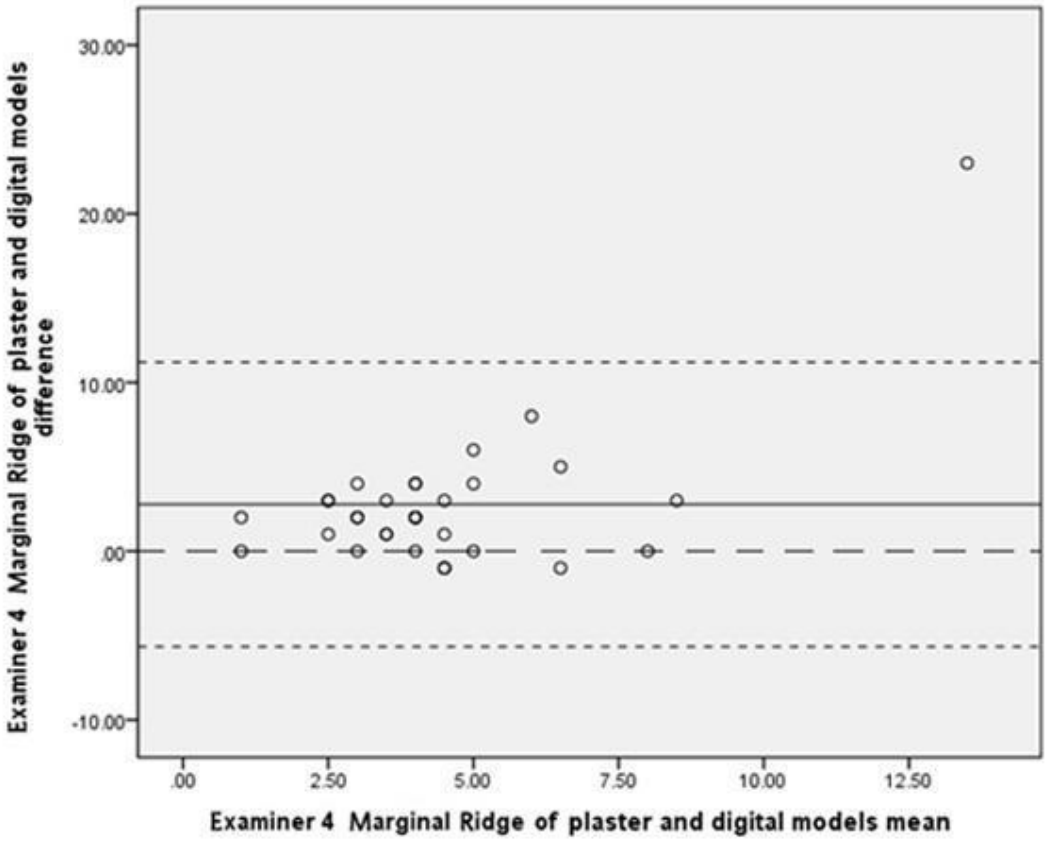


Figure 5.28 Bland-Altman scatter plot for Marginal Ridge for Examiner 4

Table 5.25 Mean and limits of agreement of plaster and digital models for Marginal Ridge for Examiner 4

The mean difference of the Marginal Ridges for Examiner 4	2.77
The limits of agreements of Marginal Ridges for Examiner 4	11.20-(-5.66)

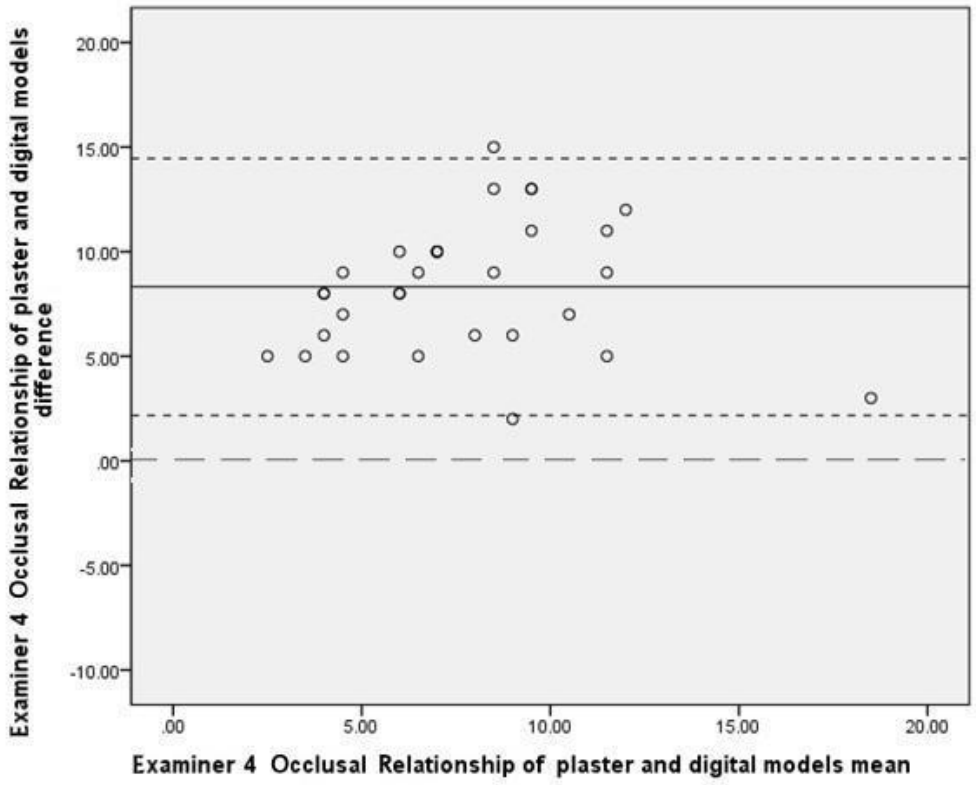


Figure 5.29 Bland-Altman scatter plot for Occlusal Relationship for Examiner 4

Table 5.26 Mean and limits of agreement of plaster and digital models for Occlusal Relationship for Examiner 4

The mean difference of the Occlusal Relationship for Examiner 4	8.32
The limits of agreements of Occlusal Relationship Examiner 4	14.46-(2.17)

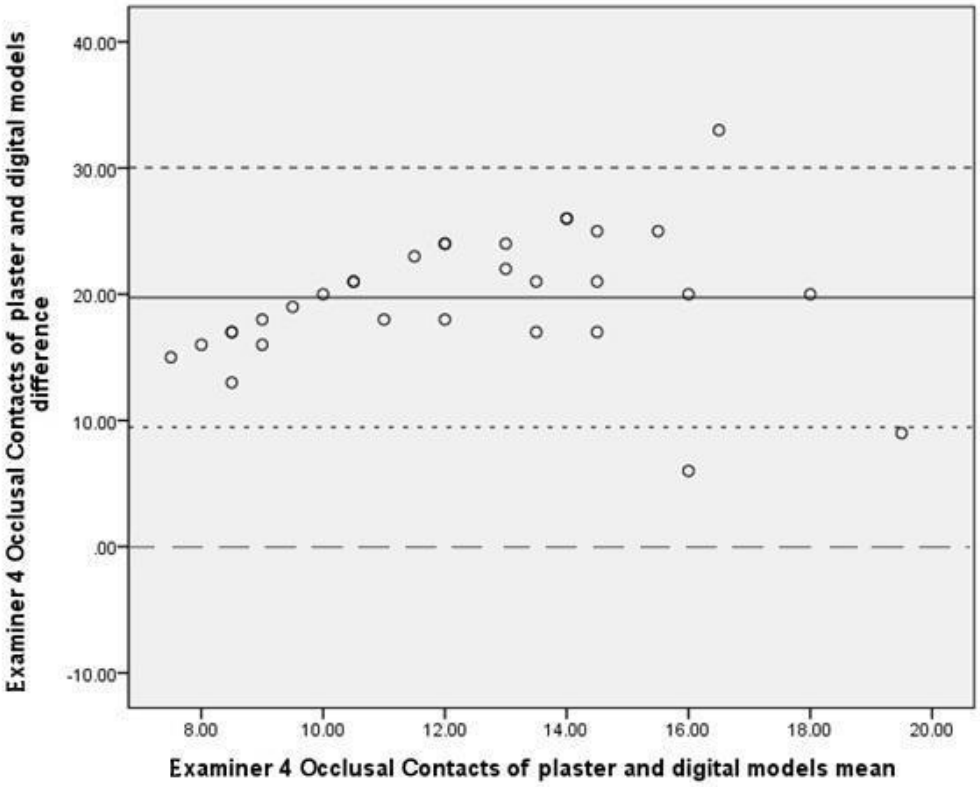


Figure 5.30 Bland- Altman scatter plot for Occlusal Contacts for Examiner 4

Table 5.27 Mean and limits of agreement of plaster and digital models for Occlusal Contacts for Examiner 4

The mean difference of the Occlusal Contacts for Examiner 4	19.74
The limits of agreements of Occlusal Contacts Examiner 4	30.026-(9.45)

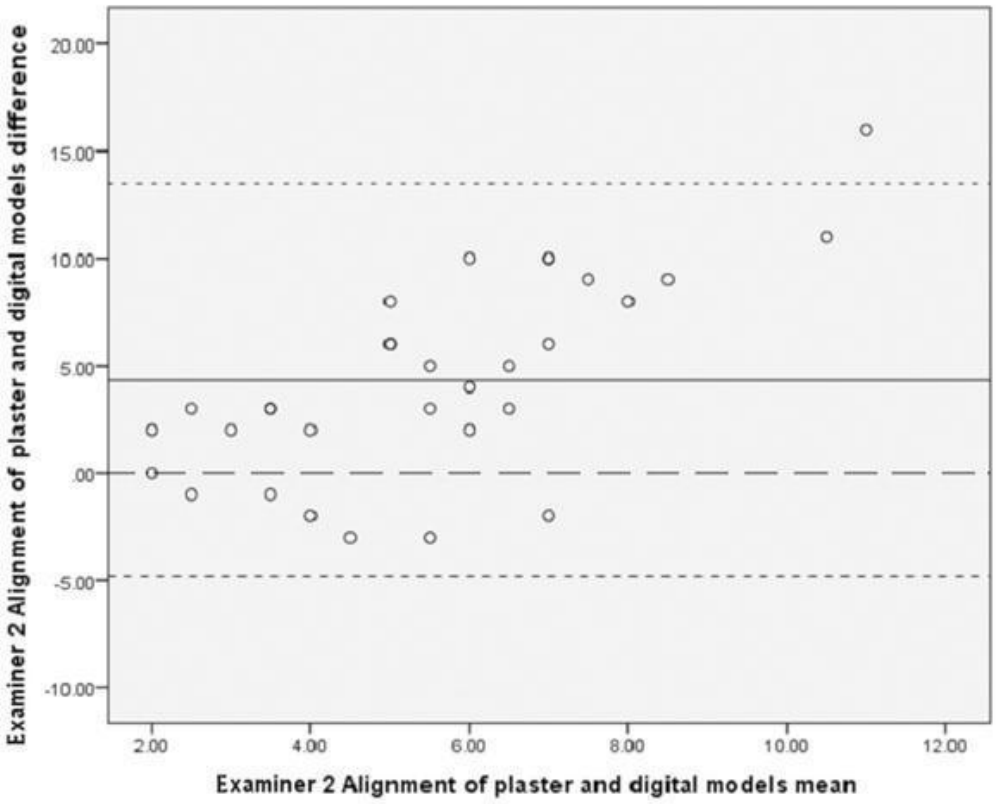


Figure 5.31 Bland-Altman scatter plot for Overjet for Examiner 4

Table 5.28 Mean and limits of agreement of plaster and digital models for Overjet for Examiner 4

The mean difference of the Overjet for Examiner 4	22.12
The limits of agreements of Overjet Examiner 4	30.60-(13.64)

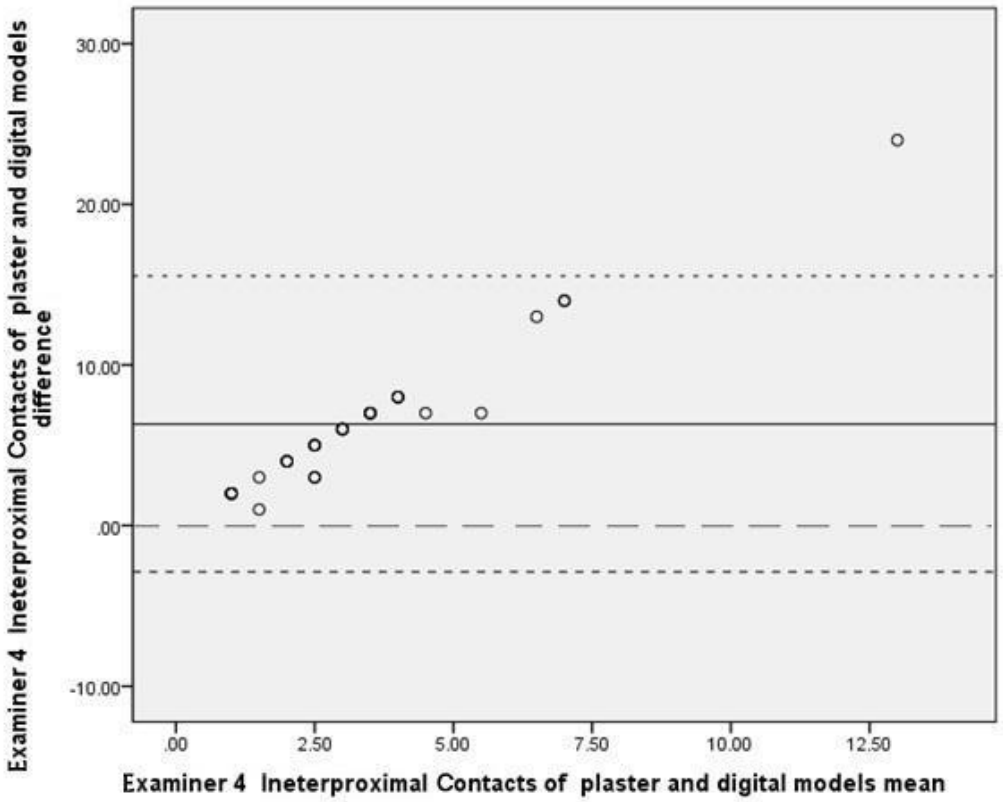


Figure 5.32 Bland-Altman scatter plot for Interproximal Contacts for Examiner 4

Table 5.29 Mean and limits of agreement of plaster and digital models for Interproximal Contacts for Examiner 4

The mean difference of the Interproximal Contacts for Examiner 4	6.32
The limits of agreements of Interproximal Contacts Examiner 4	15.53 –(-2.89)

Appendix 2: Grading System for Dental Casts and Panoramic Radiographs

The American Board of Orthodontics

Grading System for Dental Casts and Panoramic Radiographs

Revised June 2012

INTRODUCTION

The American Board of Orthodontics is constantly striving to make the clinical examination a fair, accurate, and meaningful experience for examinees. In an effort to enhance their liability of the examiners and provide the examinees with a tool to assess the adequacy of their finished orthodontic results, the Board has established a Model Grading System to evaluate the final dental casts and panoramic radiographs. This scoring system was developed systematically through a series of four field tests over a period of five years. The Board instituted the model and radiographic portions of the Model Grading System, and it has been used to grade these portions of the examinees' clinicalcasereportssince1999. In an effort to assist examinees with the selection of their cases, the Board is making this Model Grading System available to all examinees. The Board encourages examinees to score their own case reports with this scoring system to determine if they meet Board standards.

BACKGROUND

In 1994, The American Board of Orthodontics began investigating methods of making the clinical examination more objective. Since a major emphasis has always been placed on the final occlusion, the first efforts were directed at developing an objective method of evaluating the dental casts and intraoral radiographs.

In the past, several indices have been used to evaluate the outcome of orthodontic treatment.[1,2,3,4] Generally, these indices compare pretreatment and posttreatment records to determine the quality of the final result. However, these indices are not precise, and the validity and reliability of these indices has not been established. The Occlusal Index[5] has also been used to determine treatment quality. However, this method is tedious, and the system is more appropriate for scoring pretreatment rather than posttreatment records.

In 1987, the PAR Index[6] (Peer Assessment Rating) was developed to assess an occlusion at any stage of development. Over 200 dental casts representing various pretreatment and posttreatment stages of occlusion were used to establish this index. The PAR Index has good reliability and validity; however

this measuring system is not precise enough to discriminate between the minor inadequacies of tooth position that are found in ABO case reports. Therefore, an ABO committee was formed in 1994, to begin field testing precise methods of objectively evaluating posttreatment dental casts and panoramic radiographs.

At the 1995 ABO clinical examination, 100 cases were evaluated. A series of 15 criteria were measured on each of the final dental casts and panoramic radiographs. The data showed that 85% of the inadequacies in the final results occurred in seven of the 15 criteria (alignment, marginal ridges, buccolingual inclination, overjet, occlusal relationships, occlusal contacts, root angulation).

Therefore, at the 1996 clinical examination, a second field-test was initiated to verify the results of the previous test and to determine if multiple examiners could score the records reliably and consistently. In this field test, a subcommittee of four Directors evaluated 300 sets of post-treatment dental casts and panoramic radiographs. Again, the majority of the inadequacies in the final results occurred in the same seven categories, but the committee had difficulty establishing adequate inter-examiner reliability. The subcommittee recommended that a measuring instrument be developed to make the measuring process more reliable.

In 1997, a third field test was performed using the modified scoring system with the addition of an instrument to measure the various criteria more accurately. All of the Directors participated in this field test, and a total of 832 dental casts and panoramic radiographs were measured. The same seven criteria were evaluated. A calibration session preceded the examination to establish more accurate use of the measuring instrument and improve the reliability of the Directors. The results again showed that the overwhelming majority of the inadequacies in the finished results occurred in the aforementioned categories. However, the Directors decided to add interproximal contacts to the scoring system to raise the total number of criteria to eight. In addition, modifications were made in the measuring instrument to improve measuring accuracy among Directors.

In 1998, the fourth and final field test was initiated. Again all Directors participated in the evaluation process. The new and improved measuring

instrument was used. An extensive training and calibration session was performed prior to the actual examination. The major objectives of this final field test were to refine the measuring and calibration process, and to gather enough data on general performance to establish the validity or cut-off for passing this portion of the clinical examination. This field test was extremely successful. Not only did it reaffirm the benefits of using an objective system for grading the dental casts and panoramic radiographs, but also it helped to establish standards for successful completion of this portion of the clinical examination.

Based upon the collective and cumulative results of these extensive field tests, the Board decided to officially initiate the use of this Model Grading System for examinees at the February 1999, ABO clinical examination in St. Louis. In order to assist the examinee in selecting cases that will successfully pass the examination process, the Board is providing the examinee with the same system used by the Directors. The Board encourages examinees to score their own dental casts and panoramic radiographs during their preparation for the clinical examination in order to select cases that will successfully pass the ABO Model Grading System.

CRITERIA AND RATIONALE

The ABO Model Grading System for scoring dental casts and panoramic radiographs contains eight criteria. These are: alignment, marginal ridges, buccolingual inclination, occlusal relationships, occlusal contacts, overjet, interproximal contacts, and root angulation. The rationale for using these criteria is stated in the following section.

Alignment is usually a fundamental objective of any orthodontic treatment plan. Therefore, it seems reasonable that any assessment of quality of orthodontic result must contain an assessment of tooth alignment. In the anterior region, the incisal edges and lingual surfaces of the maxillary anterior teeth and the incisal edges and labial- incisal surfaces of the mandibular anterior teeth were chosen as the guide to assess anterior alignment. These are not only the functioning areas of these teeth, but they also influence esthetics if they are not arranged in proper relationship. In the maxillary posterior region, the

mesiodistal central groove of the premolars and molars is used to assess adequacy of alignment. In the mandibular arch, the buccal cusps of the premolars and molars are used to assess proper alignment. These areas were chosen since they represent easily identifiable points on the teeth, and represent the functioning areas of the posterior teeth. The results of the four field tests show that the most commonly mal aligned teeth were the maxillary and mandibular lateral incisors and second molars, which accounted for nearly 80% of the mistakes.

Marginal ridges are used to assess proper vertical positioning of the posterior teeth. In patients with no restorations, minimal attrition, and no periodontal bone loss, the marginal ridges of adjacent teeth should be at the same level. If the marginal ridges are at the same relative height, the cementoenamel junctions will be at the same level. In a periodontally healthy individual, this will result in flat bone level between adjacent teeth. In addition, if marginal ridges are at the same height, it will be easier to establish proper occlusal contacts, since some marginal ridges provide contact areas for opposing cusps. Based upon the four field tests, the most common mistakes in marginal ridge alignment occurred between the maxillary first and second molars. The second most common problem area was between the mandibular first and second molars.

Buccolingual inclination is used to assess the buccolingual angulation of the posterior teeth. In order to establish proper occlusion in maximum intercuspation and avoid balancing interferences, there should not be a significant difference between the heights of the buccal and lingual cusps of the maxillary and mandibular molars and premolars. The Directors use a special step gauge to assess this relationship. Some latitude is allowed, however in past field tests significant problems were observed in the buccolingual inclination of the maxillary and mandibular second molars.

Occlusal contacts are measured to assess the adequacy of the posterior occlusion. Again, a major objective of orthodontic treatment is to establish maximum intercuspation of opposing teeth. Therefore, the functioning cusps are used to assess the adequacy of this criterion; i.e., the buccal cusps of the mandibular molars and premolars, and the lingual cusps of the maxillary molars

and premolars. If cusp form is small or diminutive, that cusp is not scored. In past field tests, the most common problem area has been inadequate contact between maxillary and mandibular second molars.

Occlusal relationship is used to assess the relative anteroposterior position of the maxillary and mandibular posterior teeth. In order to achieve accuracy and reliability in measuring this relationship, results of previous field tests have shown that the most verifiable method of scoring this criterion is to use Angle's relationship. Therefore, the buccal cusps of the maxillary molars, premolars, and canines must align within 1 mm of the interproximal embrasures of the mandibular posterior teeth. The mesiobuccal cusp of the maxillary first molar must align within 1 mm of the buccal groove of the mandibular first molar.

Overjet is used to assess the relative transverse relationship of the posterior teeth, and the anteroposterior relationship of the anterior teeth. In the posterior region, the mandibular buccal cusps and maxillary lingual cusps are used to determine proper position within the fossae of the opposing arch. In the anterior region, the mandibular incisal edges should be in contact with the lingual surfaces of the maxillary anterior teeth. In past field tests, the common mistakes in overjet have occurred between the maxillary and mandibular incisors and second molars.

Interproximal contacts are used to determine if all spaces within the dental arch have been closed. Persistent spaces between teeth after orthodontic therapy are not only unesthetic, but can lead to food impaction. In past field tests, spacing is generally not a major problem with ABO cases.

Root angulation is used to assess how well the roots of the teeth have been positioned relative to one another. Other than periapical radiographs or three-dimensional imaging, the panoramic radiograph is probably the best practical means for making this assessment. It is incumbent upon the examinee to present imaging evidence to document post-treatment root position. If roots are properly angulated, then sufficient bone will be present between adjacent roots, which could be important if the patient were susceptible to periodontal bone loss at some point in time. If roots are dilacerated, then they are not graded. In past

field tests, the common mistakes in root angulation occurred in the maxillary lateral incisors, canines, second premolars, and mandibular first premolars.

GUIDE FOR GRADING CLINICAL CASE REPORTS MODEL ANALYSIS

ALIGNMENT

In the maxillary and mandibular anterior regions, proper alignment is characterized by coordination of alignment of the incisal edges and lingual incisal surfaces of the maxillary incisors and canines(fig.1),and the incisal edges and labial incisal surfaces of the mandibular incisors and canines(fig.2).

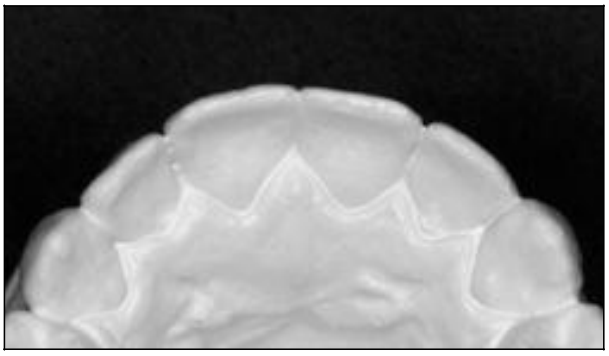


Figure 1

Figure 2

In the mandibular posterior quadrants, the mesiobuccal and distobuccal cusps of the molars and premolars should be in the same mesiodistal alignment. In the maxillary arch, the central grooves (mesio-distal) should all be in the same plane or alignment (fig. 3). If all teeth are in alignment, or within 0.50 mm of proper alignment, no points are scored.

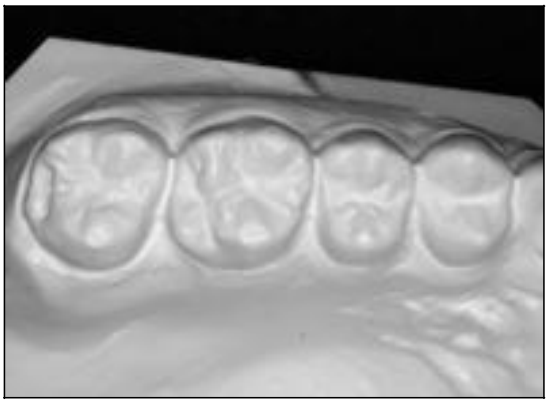


Figure 3

If the mesial or distal alignment at any of the contact points is 0.50 mm to 1 mm deviated from proper alignment (fig. 4a,b), 1 point shall be scored for the tooth that is out of alignment. If adjacent teeth are out of alignment, then 1 point should be scored for each tooth.

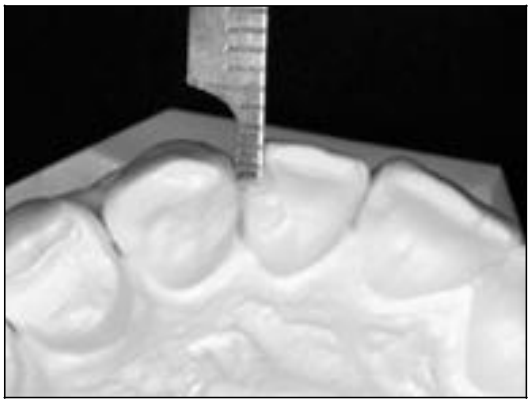


Figure 4a

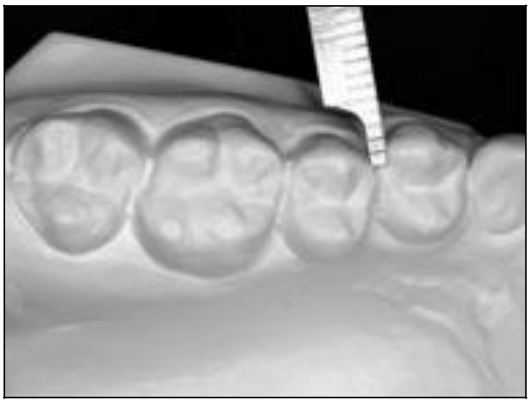


Figure 4b

If the discrepancy in alignment of a tooth at the contact point is greater than 1 mm, then2 points shall be scored for that tooth (fig. 5a,b). No more than 2 points shall be scored for any tooth.

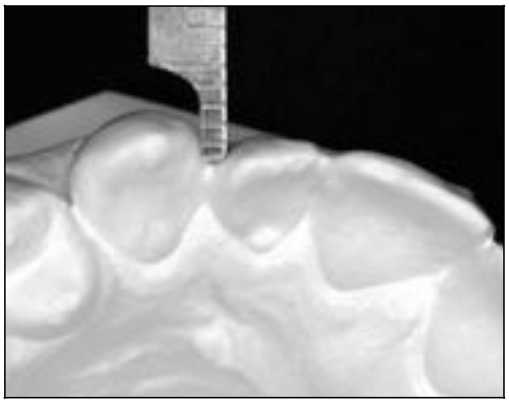


Figure 5a

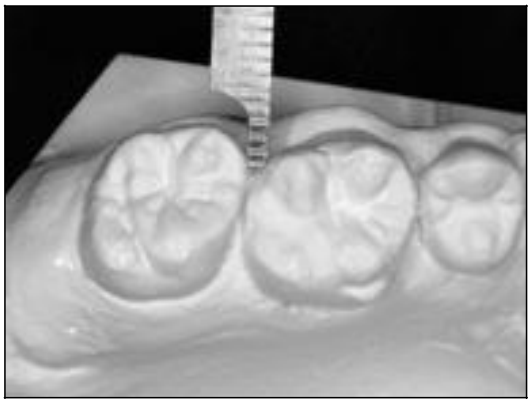


Figure 5b

MARGINAL RIDGES

In both maxillary and mandibular arches, marginal ridges of adjacent posterior teeth shall be at the same level, or within 0.50 mm of the same level (fig. 6).

Figure 6

In scoring, do not include the canine-premolar contact; and do not include the distal of lower 1st premolar.

If adjacent marginal ridges deviate from 0.50 to 1 mm (fig. 7), then 1 point is scored for that interproximal contact. If the marginal ridge discrepancy is greater than 1 mm (fig.8), then 2 points shall be scored for that interproximal contact. No more than 2 points will be scored for any contact point. The marginal ridge will be considered as the most occlusal point that is within 1 mm of the contact at the occlusal surface of adjacent teeth.

Figure 7

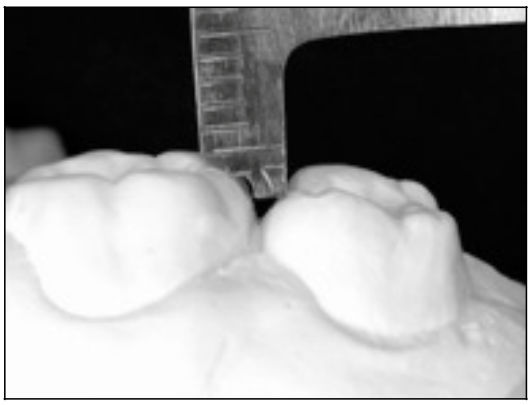


Figure 8

BUCCOLINGUAL INCLINATION

The buccolingual inclination of the maxillary and mandibular posterior teeth shall be assessed by using a flat surface that is extended between the occlusal surfaces of the right and left posterior teeth. When positioned in this manner, the straight edge should contact the buccal cusps of contralateral mandibular molars and premolars. The lingual cuspsshouldbewithin1mmofthe surface of the straight edge (fig. 9). In the maxillary arch, the straight edge should contact the lingual cusps of the maxillary molars and premolars. Thebuccalcuspsshouldbewithin1 mm of the surface of the straight edge (fig.10).

Figure 9

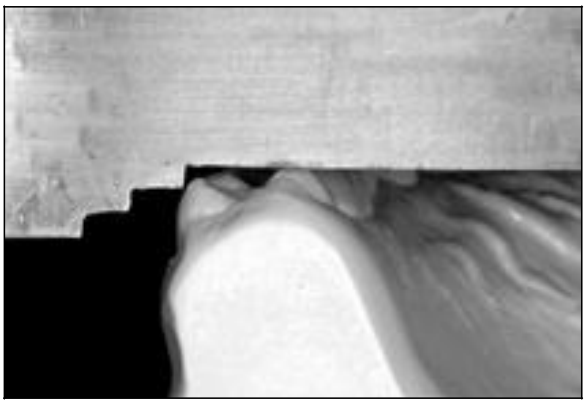


Figure 10

Do **not** score the mandibular 1st premolars nor the distal cusps of the second molars.

If the mandibular lingual cusps or maxillary buccal cusps are more than 1 mm, but less than 2 mm from the straight edge surface (fig. 11a,b), 1 point shall be scored for that tooth.

Figure 11a

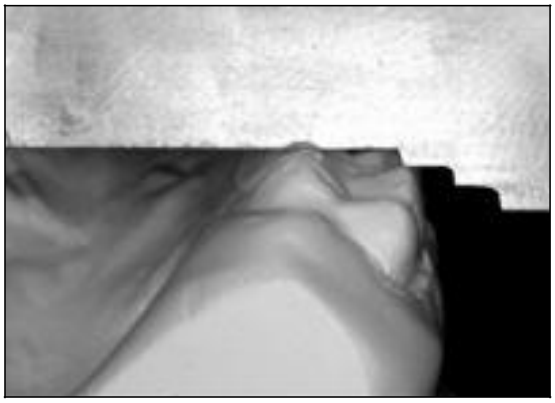


Figure 11b

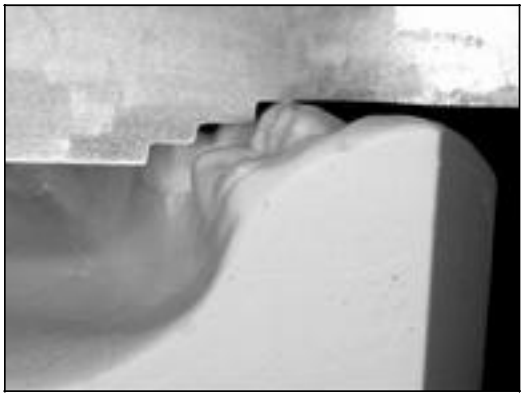


Figure 12a

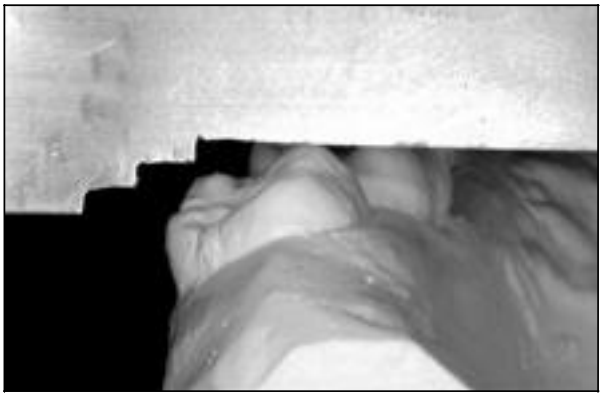


Figure 12b

If the discrepancy is greater than 2 mm (fig. 12a, b), then 2 points are scored for that tooth. No more than 2 points shall be scored for any tooth.

OCCLUSAL CONTACTS

This section of the evaluation determines the adequacy of occlusal contact of the premolars and molars. The buccal cusps of the mandibular premolars and molars (fig.13) and the lingual cusps of the maxillary premolars and molars (fig. 14) should be contacting the occlusal surfaces of the opposing teeth. Each mandibular premolar has one functional cusp. Each mandibular molar has two functional buccal cusps. The maxillary premolars have one functional lingual cusp. However, the maxillary molars may have only a mesiolingual functional cusp.

Figure 13

Figure 14

If the distolingual cusp is short or diminutive (fig. 15), it should not be considered in the evaluation. If this cusp is prominent, but does not contact with the opposing arch, then points may be scored. If the cusps are in contact with the opposing arch, no points are scored. Do not score diminutive

distolingual cusps of the maxillary 1st and2nd molars, nor lingual cusps of the mandibular first premolars.

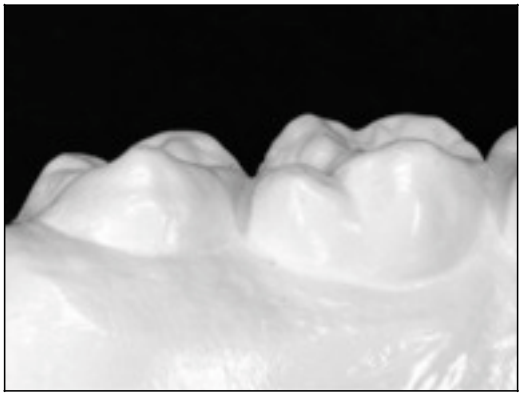


Figure 15

If a cusp is out of contact with the opposing arch, and the distance is 1 mm or less (fig.16), then 1 point is scored for that tooth. If the cusp is out of contact and the distance is greater than 1 mm (fig. 17), then 2 points are scored for that tooth. No more than 2 points are scored for each tooth.

Figure 16

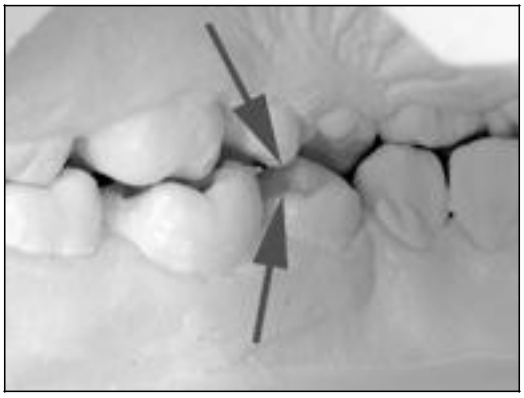


Figure 17

OCCLUSAL RELATIONSHIP

This section of the evaluation determines whether the occlusion has been finished in an Angle Class I relationship. Ideally, the maxillary canine cusp tip should align with (or within 1 mm of) the embrasure or contact between the mandibular canine and adjacent premolar (fig. 18). The buccal cusps of the maxillary premolars should align with (or be within 1 mm of) the embrasures or contacts between the mandibular premolars and first molar (fig. 18). The mesiobuccal cusps of the maxillary molars should align with (or be within 1 mm of) the buccal grooves of the mandibular molars (fig. 18).

Figure 18

If the maxillary buccal cusps deviate between 1 and 2 mm from the aforementioned positions (fig. 19), then 1 point shall be scored for that maxillary tooth. If the buccal cusps of the maxillary premolars or molars deviate by more than 2 mm from ideal position (fig. 20), then 2 points shall be scored for each maxillary tooth that deviates. No more than 2 points shall be scored for each maxillary tooth. In some situations, the posterior occlusion may be finished in either an Angle Class II or Class III relationship, depending upon the type of tooth extraction in the maxillary or mandibular arches.

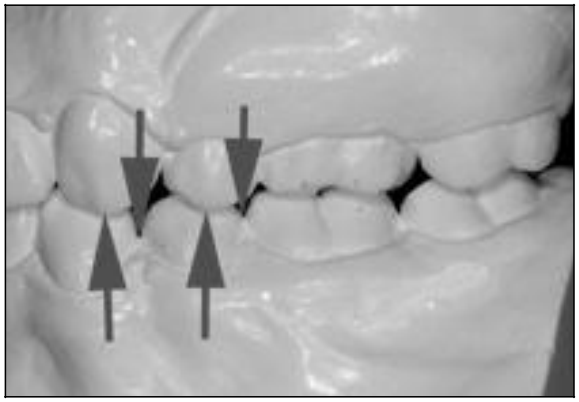


Figure 19

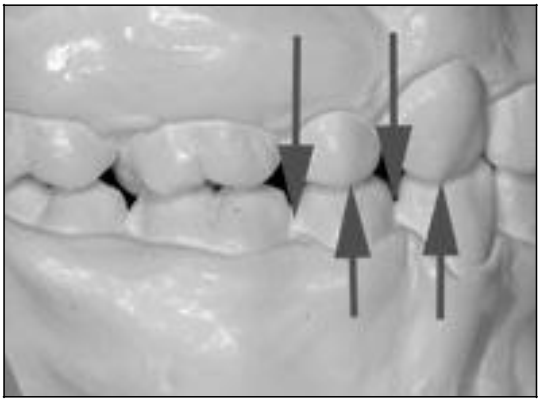


Figure 20

In a Class II situation (fig. 21), the buccal cusp of the maxillary first molar should align with the embrasure or interproximal contact between the mandibular second premolar and first molar. The buccal cusp of the maxillary second molar should align with the embrasure or interproximal contact between the mandibular first and second molars. If the final occlusion is finished in a Class III relationship (when mandibular premolars are extracted), the buccal cusp of the maxillary second premolar should align with the buccal groove of

the mandibular first molar (fig. 22). The remaining occlusion distal to the maxillary second premolar and mandibular first molar are adjusted accordingly.

Figure 21

Figure 22

OVERJET

The overjet is evaluated by articulating the models and viewing the labiolingual relationship of the maxillary arch relative to the mandibular arch. In order to determine the proper relationship of the casts, the examiner must rely on the trimming of the backs of the bases of the models. The models are set flat on their backs, in order to determine this assessment (fig. 23).

Figure 23

If the models are mounted on an articulator, then the articulated mounting shall determine the proper maxillary and mandibular model relationship. If the proper overjet has been established, then the buccal cusps of the mandibular molars and premolars will contact in the center of the occlusal surfaces, buccolingually, of the maxillary premolars and molars (fig. 24). In the anterior region, the mandibular canines and incisors will contact the lingual surfaces of

the maxillary canines and incisors (fig. 25). If this relationship exists, no points are scored.

Figure 24

Figure 25

If the mandibular buccal cusps deviate 1 mm or less from the center of the opposing tooth (fig. 26), 1 point is scored for that tooth. If the position of the mandibular buccal cusps deviates more than 1 mm from the center of the opposing tooth (fig. 27), two points are scored for that tooth. No more than 2 points are scored for any tooth.

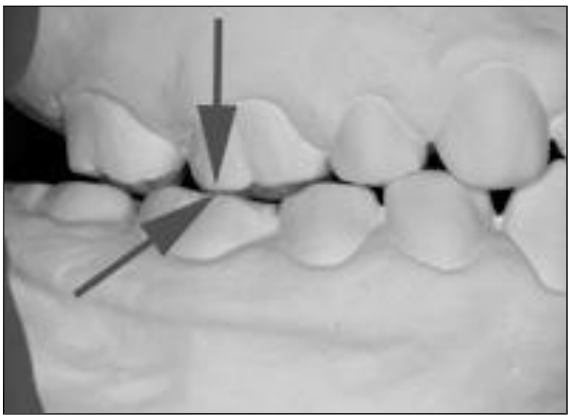


Figure 26

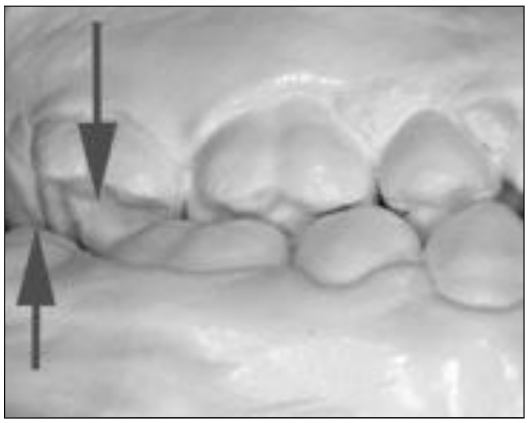


Figure 27

In the anterior region, if the mandibular canines or incisors are not contacting lingual surfaces of the maxillary canines and incisors, and the distance is 1 mm or less (fig. 28), then 1 point is scored for each maxillary tooth. If the discrepancy is greater than 1 mm (fig. 29), then 2 points are scored for each maxillary tooth.

Figure 28

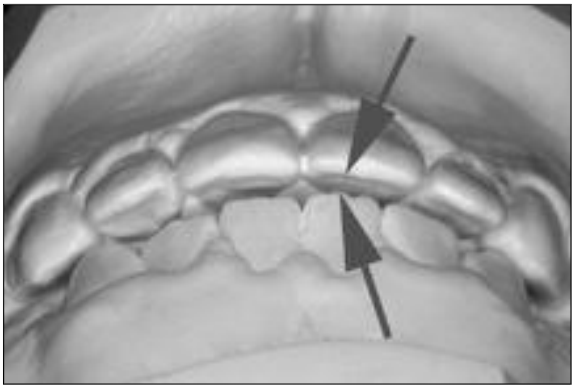


Figure 29

Note that although Overjet is typically scored by assessing contact between opposing teeth, this score is subject to examiner modification. For example, cases in which incisors display extremely acute inter-incisal angles and/or significant overlap of incisal edges may be scored an additional point.

INTERPROXIMAL CONTACTS

This assessment is made by viewing the maxillary and mandibular dental casts from an occlusal perspective. The mesial and distal surfaces of the teeth should be in contact with one another (fig. 30). If 0.50 mm or less interproximal space exists, then no points are scored.

Figure 30

If greater than 0.50 to 1 mm of interproximal space exists between two adjacent teeth (fig. 31), then 1 point is scored for that interproximal contact. If more than 1 mm of space is present between two teeth (fig. 32), then 2 points are scored for that interproximal contact. No more than 2 points are scored for any contact that deviates from ideal.

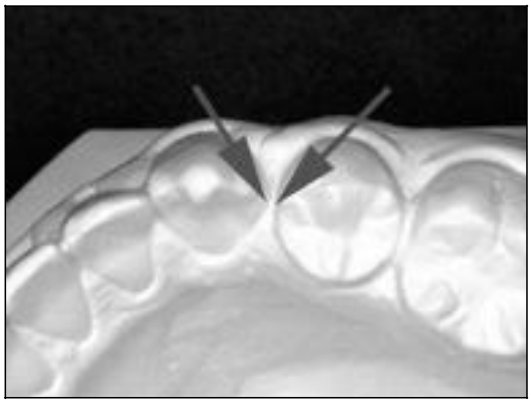


Figure 31

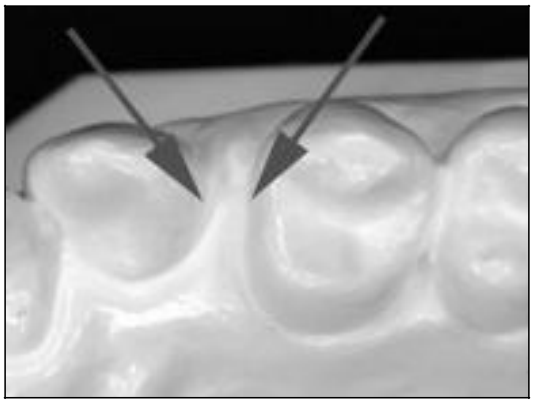


Figure 32

RADIOGRAPHIC ANALYSIS

ROOT ANGULATION

The relative angulation of the roots of the maxillary and mandibular teeth is assessed on the panoramic radiograph. Although this is not ideal, it gives a reasonably good assessment of root position. Generally, the roots of the maxillary and mandibular teeth should be parallel to one another and oriented perpendicular to the occlusal plane (fig.33). If this situation exists, then no points are scored.

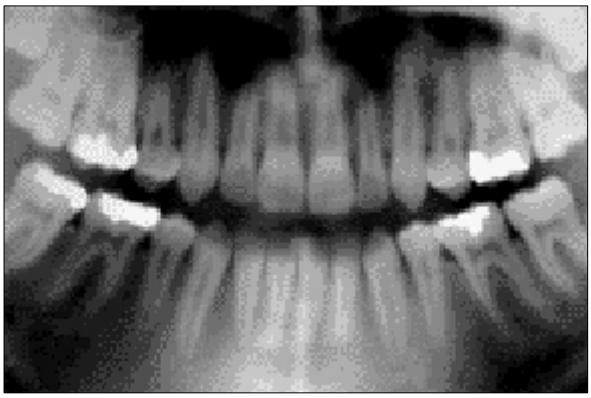


Figure 33

The ABO acknowledges the distortion that frequently occurs within panoramic radiographs. The Board has recommended the following:

Omit scoring the canine relationship with adjacent tooth root when using a final panoramic radiograph.

If a root is angled to the mesial or distal (not parallel) and is close to, but not touching, the adjacent tooth root, then 1 point is scored for each discrepancy (anterior, premolar, and/or molar areas, fig. 34). If the root is angled to the mesial or distal and is contacting the adjacent tooth root (fig. 35), then 2 points are scored for that tooth.

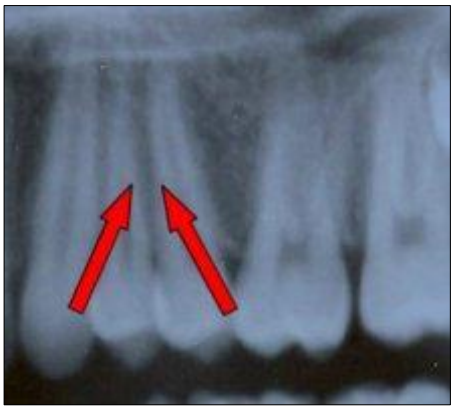


Figure 34

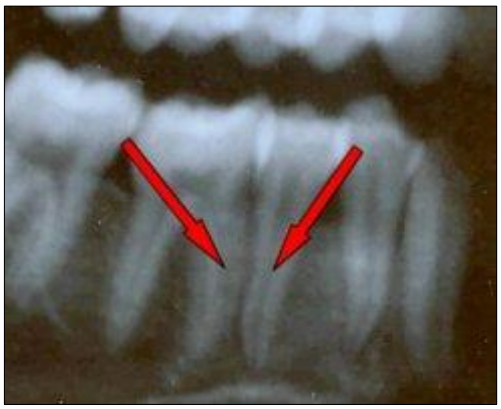


Figure 35

EVALUATION OF CASES

The Board's decision to evaluate an individual case as Complete or Incomplete is based upon multiple factors. Record quality and the ability to finish a case are important, but they are not the only aspects that are considered in the evaluation. Case management,a sound understanding of diagnosis, treatment planning and mechanotherapy are equally important and are discussed during the actual interview when cases are reviewed with the examinee.

A score corresponding to Complete in the Cast-Radiograph Evaluation and Case Management are determined at every clinical examination during a pre-exam calibration session of all examiners. Therefore, scores for cases evaluated as Complete will vary from exam to exam and may range from:

- 27 or less for C-R Eval
- 7 or less for CMF
- And, case meets DI and case criteria

High scores on individual segments, or combinations of individual segments, may cause a case to become Incomplete. From time to time, however, a successful interview may result in an overturn of an otherwise Incomplete case.

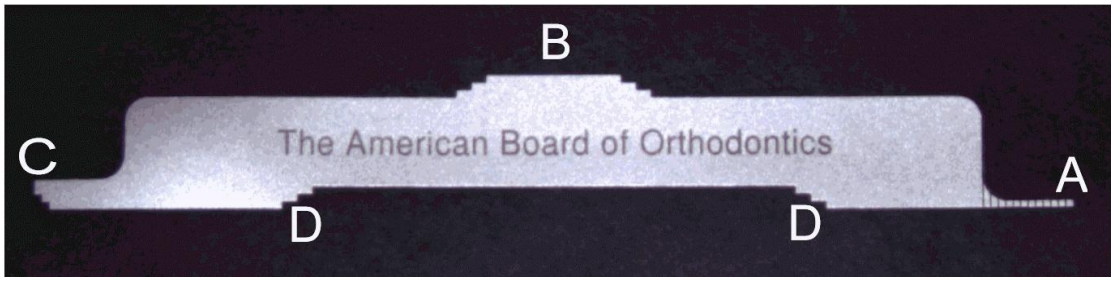
SUMMARY

The Directors of The American Board of Orthodontics have spent countless hours developing this system for assessing the occlusal and radiographic results of orthodontic treatment. The usefulness of this system depends not only on its objectivity, but more importantly on the validity and reliability of the measurements. After repeated comparison of both objective and subjective systems, the Directors are confident that the "cut-off" score to pass this portion of the clinical examination is valid. Reliability will be insured through the use of a precise measuring instrument, in addition to training and calibration of the Directors before each examination. In order to be fair to all examinees, a confidence interval is established to account for interrater variability.

Although the underlying purpose of establishing this grading system is to insure reliable, objective evaluation of orthodontic records, the Board sees a much greater benefit to publishing this grading system. Now, examinees may grade their own results before the clinical examination and know if their results will

pass Board standards. Furthermore, Diplomates may use this scoring system at anytime in their clinical career to determine if they are producing "Board quality" results. The Board hopes that this method of self- evaluation will help to elevate the overall quality of orthodontic care.

ABO MEASURING GAUGE



A This portion of the gauge is in 1 mm increments and is used to measure discrepancies in alignment, overjet, occlusal contact, interproximal contact, and occlusal relationships. The width of the gauge is 0.5 mm.

B This portion of the gauge has steps measuring 1 mm in height and is used to determine discrepancies in mandibular posterior buccolingual inclination.

C This portion of the gauge has steps measuring 1 mm in height and is used to determine discrepancies in marginal ridges.

D This portion of the gauge has steps measuring 1 mm in height and is used to determine discrepancies in maxillary posterior buccolingual inclination.

NOTE: Third molars are not scored unless they substitute for the second molars.

You may download the ABO Grading System for Casts-Radiographs from the ABO website>Orthodontic Professionals > Clinical Examination > Download and Print: Forms and References.

This gauge is included in the Calibration Kit along with three sets of pre measured cases. There is a digital component to the Calibration Kit which arrives as an attachment to the email receipt of purchase. The digital component contains the grading system manual, panoramic radiographs and scoring keys.

REFERENCES

1. Eismann, D A method of evaluating efficiency of orthodontic treatment, Trans EuropOrthodSoc, 223-232, 1974
2. Eismann, D Reliable assessment of morphological changes resulting from orthodontic treatment, Europ J Orthod, 2:19-25, 1980
3. Gottlieb, E Grading your orthodontic treatment results, J ClinOrthod, 9:156-161, 1975
4. Berg, R Post-retention analysis of treatment problems and failures in 264 consecutively treated cases, Europ J Orthod, 1:55-68, 1979
5. Summers, C The occlusal index: a system for identifying and scoring occlusal disorders, Am J Orthod, 59:552-566, 1971
6. Richmond,S.,Shaw,W,etal.ThedevelopmentofthePARIndex(PeerAssessmentRating):reliabilityandvalidity,EuropJOrthod,14:125-139,1992
7. Mckee, I., Williamson, C., et al. The accuracy of 4 panoramic units in theprojection of mesiodistal tooth angulations, AJO-DO 2002, 121:166-175
8. Peck,J.,Sameshima,G.,etal.MesiodistalRootAngulationUsingPanoramicandConeBeamCT,AngleOrthodontist2007,No.2:206-213
9. Owens,A.M.,Johal,A.NearEndofTreatmentPanoramicRadiographintheAssessmentofMesiodistalRootAngulation,AngleOrthodontists,Vol78,No.3:475-481